RiCL Research in Corpus Linguistics

Corpus as a slice of life: Representing naturally occurring language and its speakers

Giorgia Troiani[a/b] – John W. Du Bois[b] – Andrey Filchenko[a]

Nazarbayev University[a] / Kazakhstan
University of California, Santa Barbara[b] / United States

Abstract – Discourse is subject to numerous forces that shape its form. One force that is underestimated is the interactional dynamic among interlocutors. In devising the criteria that inform data selection for a corpus of spoken discourse, designers may end up prioritizing the collection of spontaneous discourse and overlook the fact that this type of discourse can still display artificial interactional dynamics. We propose an approach to spoken corpus compilation that aims at preserving naturally occurring interactional dynamics by choosing as focus of the corpus the representation of participants' lives. Through the analysis of speech events collected in different projects, we demonstrate the advantages of sourcing naturally occurring discourse over spontaneous data. We then discuss a series of practices that the authors implemented in different contexts to ensure the collection of naturally occurring data. We argue that this framework yields the construction of corpora that are representative not only of a language, but also of the lives of its users.

Keywords – corpus design; naturally occurring discourse; spontaneous discourse; recording practices

1. INTRODUCTION[1]

When creating a corpus of everyday spoken discourse, designers must balance their theoretical assumptions about language with time and resource constraints. They do so by establishing strict criteria that guide the decision over which data to include in the corpus. One productive approach is to prioritize informal spontaneous discourse (e.g., conversation), either exclusively (Chui and Lai 2008; Raso and Mello 2012; Love *et al.* 2017; Goria and Mauri 2018) or in combination with other genres (Greenbaum 1991; Burnard 2002; Kucera 2002; Oostdijk 2002). While this approach guarantees the collection of everyday conversational data, it places the corpus designers' attention on

---

structural features of language, promoting its treatment as a series of constructions disembodied from the interactional context in which they are produced. An alternative to this approach is the 'cast the wide net' model (Du Bois and Troiani 2022), a framework that integrates corpus linguistics with anthropology, first implemented in the design of the *Santa Barbara Corpus of Spoken American English* (SBCSAE; Du Bois *et al.* 2000). Within this framework, the focus of corpus design is shifted from the representation of language to the representation of the participants' behavior and their lives.[2]

In this paper, we present the theoretical foundations that underpin the 'cast the net wide' framework. The second author originally developed the framework in the construction of the SBCSAE. The first and third authors implemented and adapted this framework in the compiling of the *Multimedia Corpus of Modern Spoken Kazakh Language* (MULTICORSKL; Filchenko *et al.* 2023), the first spoken corpus of Kazakh. Our experience with corpus construction has been informed by disciplines like anthropology and language documentation. Drawing on the discourse and conversational analysis of different types of speech events, we illustrate an approach to spoken corpus design that prioritizes the role of individual speech events in participants' lives as criterion for data collection. We demonstrate how speech events that are selected according to this criterion (i.e., naturally occurring data) can differ from events that prioritize different criteria. We particularly stress how 'spontaneous' data are not necessarily naturally occurring. We argue that naturally occurring data preserve interactional social dynamics between participants to a speech event, which in turn results in the construction of a corpus which aims to be representative of the lives of participants, rather than of linguistic structures. In the rest of this section, we show how spontaneous and naturally occurring types of discourse differ from each other by analyzing the interactional dynamics of similar speech events collected with different protocols. We then illustrate the shortcomings with restricting data selection to spontaneous discourse (Section 3) and the adjustments introduced to our data collection protocol to ensure the collection of naturally occurring discourse (Section 4). We base this paper on the analysis of data from the MULTICORSKL and of recordings produced for previous projects (Section 2).

Within the 'cast the net wide' framework, the goal of a corpus is to capture the context in which language arises, i.e., the life of participants, rather than language itself.

---

[2] While our expertise is on spoken discourse, many of these considerations are valid for sign languages as well.

We can visualize these two different approaches in metaphorical terms by considering archaeological artifacts. If one is interested in Roman art, they can observe it in a museum, where curators have selected representative pieces, reconstructed their cultural functions, and established connections for the visitors. Alternatively, they can visit the archaeological site of Pompeii, a southern Italian town destroyed by a volcanic eruption in 79 AD, walk through perfectly preserved neighborhoods, and observe an artifact in the spatial and cultural context where it originated. In this way, one can notice that Roman mosaics are found not on walls, but on the floors of busy areas (pools and halls), and one may deduce that they were not intended to be used solely as decorations, but also as means of protecting floors and supporting foot traffic. Once the function of Roman mosaics is clear, it becomes evident why they are made of sturdy stone tesserae and lack the color vibrancy of their Byzantine counterparts, which were produced mainly as decorative pieces and crafted from glazed glass.

In a similar way, corpora can represent linguistic structures (corpus as exhibition) or the life events in which structures are used (corpus as archaeological site). Depending on the goal that is prioritized, the same speech event (e.g., a conversation) can end up looking different. As discussed above, we present an approach to corpus construction named 'cast the wide net' which yields naturally occurring discourse, i.e., discourse produced by the participants to fulfill life needs that are independent of the researcher's agenda. Providing examples sourced using different data collection protocols we show how naturally occurring data differ from elicited discourse, i.e., discourse prompted by a researcher, and does not overlap with spontaneous discourse, i.e., discourse that was not planned before the moment of production.

The consequences of data collection approaches on the linguistic structures of an event can be observed in Text 1 that showcases distribution and use of backchannel and features a naturally occurring and spontaneous conversation between two spouses. Backchannels are verbal or non-verbal messages used to acknowledge that the interlocutor holds the floor, and that the interaction can proceed (Drummond and Hopper 1993; Heinz 2003). Different languages exhibit different forms of backchannel.

TEXT 1

1	ADILET;[3]	*jeŋgem qaytıs bolğanda otıramın dedi me?*	Did my sister-in-law say she will sit when she is dead?
2		*Birdeŋe dep söylep.*	She said something.
3		*Endi bar,*	Anyway,
4		*Qanatqa ayt degen şıqtı Qanatqa bilmeymin?*	To Qanat – she said – "don't I know that they went to Qanat?"
5		*Birew aparmasa.*	Unless someone will catch him.
6	AISHA;	*Ïä.*	Yes.
7	ADILET;	*Endi sonımen ägi aqşanı köpirtip jür ğoy ol.*	'At the moment he is making money.'
8		*Mïlliondı.*	Millions.
9	AISHA;	*Ägi eki jüz jïr- eki mïlliondı.*	Two hundred and two millions
10	ADILET;	*Eki jüz teŋgege şıt et #,*	Two hundred tenge for fresh meat

In this excerpt, one of the speakers (Adilet) is dominating the conversation, while the other (Aisha) limits her contribution to backchannel (intonation unit (IU) 6) and occasionally provides details to co-construct a narrative (IU 9). In the text, backchannel is used with the typical function of acknowledging speakership and signaling engagement in the exchange.

Let us now compare Text 1 with a conversation (Text 2) with similar characteristics (e.g., it is spontaneous, contains the same gender dynamics, and features a similar unbalanced distribution of turns among interlocutors), but was collected through elicitation. Text 2 is sourced from a recording produced during the training phase of the MULTICORSKL. It was excluded from the corpus, as it does not feature naturally occurring discourse.

TEXT 2

129	QAIRAT;	*mağan jalpı osı*	to me in general
130		*arab tili degende*	when they [Arabic speakers] say 'Arabic'
131		*yağnï olar nege*	I mean why do they
132		*sonşama ülken*	in a territory that covers such a
133		*anday awmaqta ornalasqan jer*	large region like that
134		*nege onı bir til dep*	why do they consider it as one
135		*sanaydı desem*	language
136		*ol negizi sayasï turğıdan ğoy bılayşa aytqanda ïä*	it is basically from a political stand point so to speak right
137		*(1.2)*	
138	BOTAGOZ;	*ımm*	uhmm
139		*(0.5)*	

---

[3] All names used in the texts are pseudonyms, except for the first author and for the speakers who asked to be recognized.

140	QAIRAT;	*jalpı*	in general
141		*ne sebepti*	for what reason
142		*olar özderi sanay ma sonı*	they themselves consider
143		*bir tilde söylesemiz dep*	that they are speaking a single language

In this excerpt, where Qairat dominates the exchange, the topic of conversation was proposed by the researcher who collected the event and Botagoz does not associate with Qairat beyond classes. This suggests that Botagoz, differently than Aisha above, does not have a reason to engage with this event besides complying with a request from a researcher. The lack of a reason for engagement is reflected in the structure of the exchange. Qairat produces long turns without encountering backchannel, and even when backchannel is present, it is triggered by a long moment of silence (IU 137). Because in naturally occurring interactions speakers tend to minimize gaps and silence between turns (Pomerantz 1984; Stivers *et al.* 2009), Botagoz's silence in Text 2 is indicative of some sort of interactional trouble. In the absence of a gap between turns, the backchannel would indicate engagement with the interlocutor (Tottie 1991), but, in this case, the silence suggests that backchannel is not fulfilling this function. In fact, backchannel is here offered in reparation to the silence so as to make up for lack of engagement.

The linguistic structures featured in the two excerpts are nested within different interactional dynamics that are a consequence of the method used to collect the data. For this reason, the form, function, and position of a linguistic element (backchannel) is subject to changes. In Text 1, both interlocutors are ultimately interested in conveying and receiving information that are consequential to their lives outside of the exchange itself. For example, Aisha provides a detail to Adilet's claim that their acquaintance Qanat made money and clarifies the exact amount of money that was earned because both people associate with Qanat and knowledge of this detail may inform their future exchanges with him. In Text 2, Botagoz and Qairat seem uninterested in the exchange which was elicited by a researcher. The event was not planned beyond the prompting of the conversation from the researcher, as such, the language of the event is spontaneous, but Botagoz has no reason to engage with Qairat's opinion beyond the moment of the specific exchange. The different levels of engagement of interlocutors with the current event is visible in the linguistic and interactional structures of the event itself. In these specific cases, it is visible in the length of turns (shorter in Text 1) and in the different distribution and function of backchannel in both excerpts.

The differences between naturally and non-naturally occurring data can be noticed at the level of discourse and interactional dynamics of an event. Consider, for example, question-answer sequences in different speech events. Question-answer sequences vary as a consequence of the degree of formality of a speech event and on the number of speakers involved, as well as on the function they fulfill in the exchange (Stivers *et al.* 2010). Table 1 shows the frequency of questions and the question-answer turn distribution of different speech events. Results in Table 1 are based on the analysis of randomly selected ten-minute-long excerpts within different Kazakh speech events. The (non-)naturally occurring conversations come from the MULTCORSKL and the podcast interview comes from content created by the Kazakh-language media group *Salem Social*. All events are dialogic.[4]

	Questions per 1,000 words	Mean of words in a response turn (by any interlocutor)	Mean of words in any turn (by dominant responder)	Questioner dominance
Natural conversation	18.4	4 (SD = 3.9)	5 (SD = 5.3)	0.56
Podcast interview	14.3	21 (SD = 17.5)	22 (SD= 27.3)	0.60
Non-natural conversation	15.8	19 (SD = 30.2)	35 (SD = 20)	0.60

Table 1: Structuring of questions in speech events with different interactional expectations

We followed the coding scheme proposed by Stivers and Enfield (2010) to identify questions and answers in the excerpts. An utterance is regarded as a question if it relies on lexico-morpho-syntactic or prosodic marking (formal question) or if it sought to elicit information, confirmation, or agreement, even in the absence of formal markers (functional question). Requests for physical actions and questions inside reported speech were excluded from the analysis. For this analysis, we used a broader understanding of answers that conflates categories of 'non-answer' and 'answer' responses (Stivers and Enfield 2010: 2624). We consider an utterance to be an answer when it engages with the question as put either directly or indirectly (e.g., instances such as *I don't know*, *maybe*, requests for repetition, etc.).

The mean of words in a response turn (column 3) gives information about the average length of a turn with the function of response. This information does not distinguish between interlocutors. The mean of words in turns produced by the dominant

---

[4] The podcast interview featured three participants, but the role of interviewer is shared by two of them.

responder (column 4) gives information about the length of any turn —regardless of its function— produced by the interlocutor that produced more turns (i.e., that spoke the most). We computed the standard deviation for the previous measures: higher values correlate with exchanges where there are few long turns and many shorter turns. Finally, we computed the questioner dominance by using as numerator the number of questions asked by the participants that produced more questions and as denominator the total number of questions in the exchange (column 5). Values close to 0.5 index a dynamic where participants produce the same number of questions (e.g., a conversation), while values close to 1 index a dynamic where the production of questions is allocated to one participant (e.g., a typical interview).

The average response length in a naturally occurring conversation is strikingly lower than in the podcast interview and non-naturally occurring conversation (4 vs. 21/19 words). Moreover, the turns in the naturally occurring conversation are of a rather homogeneous length (SD = 3.9), while the turns in other events combine short turns with longer turns (SD = 17.5 for the podcast and SD = 30.2 for the non-naturally occurring conversation). This distribution is not unusual for the podcast interview, where responses can consist either of one-word answers to polar yes/no questions or of long uninterrupted sequences that are prompted by the interviewer as a mean to elaborate on the one-word answers. This behavior is maintained by the main responder throughout the entire interview: the speaker produces many one-word turns (backchannel and polar answers) and fewer long uninterrupted turns (mean = 22; SD = 27.3). Again, this is expected in an interview. In the non-naturally occurring conversation the main responder mirrors this behavior producing many one-word turns interspersed with uninterrupted turns that occasionally reach above 100 words of length (mean = 35; SD = 20). Moreover, questions tend to be slightly unidirectional, with one of the speakers asking questions more than the other (questioner dominance = 0.6), as it is the case for the podcast interview.

The behavior of the non-naturally occurring conversation can be explained by the research protocol used to collect the event. Participants to dialogic events that have been elicited do not have a reason to engage with their interlocutor other than to comply with the researcher's requests. In the absence of a reason to engage with the interlocutors, they may not have an interest in negotiating the role of speaker in the conversation.[5]

---

[5] In conversation, the visible signs of this negotiation are elements like overlap, backchannel, disfluencies, and other devices that are used to either claim or give the floor.

Interactional dynamics that distribute the conversational roles in a rigid fashion could be a convenient fallback to maintain the exchange enough to fulfill the researcher's request. In other words, one of the participants takes over the role of main questioner and directs the exchange, prompting the responder to produces responses that are quite long for a conversation. This interactional dynamic results in an exchange where both speakers produce relatively long uninterrupted turns that receive no signals of engagement from the other participant. This situation is unmatched both in the naturally occurring conversation, where the two speakers visibly engage with each other and, in the podcast, where the main responder occasionally shifts out of their role to engage the interlocutors in the co-constructions of narratives (De Fina and Perrino 2011).

This analysis shows that spontaneous and naturally occurring conversations are non-orthogonal categories and that interactional dynamics contribute to the structuring of discourse in ways that cannot be explained solely in terms of the genre of a speech event. Given this, it is worth considering whether different approaches to the collection of spoken discourse equally respect the naturally occurring interactional dynamics of an event. We argue that focusing on representing the lives of speakers, rather than language structural features alone, reduces the risk of introducing artificial dynamics to the data. For this reason, we suggest an approach to corpus compiling that prioritizes the representation of the lives of its participants. In the following sections, we demonstrate how this goal can be achieved through the 'cast the net wide' framework.

2. DATA

Data for this paper consist of recordings from different varieties collected by the authors in the context of projects with different goals and a diverse range of recording protocols. All the recordings have been transcribed into IUs according to the Discourse Functional Transcription system (Du Bois *et al.* 1993). These data are complemented by excerpts from the SBCSAE (Du Bois *et al.* 2000) and the *Switchboard Corpus* (Godfrey *et al.* 1992).

Kazakh data are extracted from the MULTCORSKL (Filchenko *et al.* 2023). As of February 2024, the corpus is composed of 150 hours of recordings (80 of which have been transcribed), produced by over 100 participants. The transcribed portion of the

corpus comprises approximately 23 thousand IUs and 600 thousand words.[6] The MULTCORSKL is a corpus of video and audio recordings of a diverse range of naturally occurring social interactions taking place in Kazakh-speaking communities in Kazakhstan and China. Among these interactions there are conversations, lectures, traditional events featuring storytelling (*aitys*), interviews, task-oriented exchanges (e.g., food preparation and games), social, and religious events. Further data will be collected and annotated through the end of 2024. We complement these data with recordings collected in the training phase of the project, which were excluded from the corpus for not fully complying with the criteria for inclusion.

Data featuring Italian and Bustocco (Western Lombard) were collected in Busto Arsizio (northern Italy) by the first author during summer 2018 as part of a documentation project. A total of three hours was collected, distributed across two conversations (one among women friends in their 70s and one between a grandmother and her nephew), an elicited narrative, and poems.

Data featuring varieties of Mixtec come from two sources. The first narrative features Jeremías Salazar speaking in Sà'án Sàvǐ ñà Yukúnanǐ, a language of the Mixtec family from Oaxaca (Mexico). The narrative was collected by the first author in the context of a collaborative documentation effort and took place in California (Salazar *et al.* 2021). The second narrative features Juan Miranda speaking in Tù'un Na Ñuu Sá Matxí Ntxè'è, another language of the Mixtec family from Oaxaca (Mexico). The event was collected in San Martin Duraznos (Oaxaca) as part of a documentation effort (Auderset and Hernández Martínez 2021). Both projects are part of a larger endeavor to document varieties of Mixtec spoken in Oaxaca, Puebla, and Guerrero (Hernández Martínez *et al.* 2021).

3. BEYOND SPONTANEOUS DISCOURSE

To meet corpus users' demand for everyday language data, one particularly productive approach employed by corpus designers is that of prioritizing the collection of spontaneous discourse; see Raso and Mello (2014) for a discussion of the reasons for the choice. Spontaneity is either identified through the analysis of the structural features of the language used in an event (Pitt *et al.* 2005) or it is assigned as a feature to specific

---

[6] Kazakh is a highly agglutinative language, so a metric like number of words is only partially informative.

genres, for example, one can collect only conversation (Love *et al.* 2017) or only events that do not display features typical of written language (Čermák 2009). In this section, we raise issues with the notion of 'spontaneity'. Firstly, we demonstrate how structurally-based definitions of spontaneity fail to capture events when participants deviate from the expected standard for their own interactional motivations. Secondly, we show how lack of spontaneity is not an inherent quality of specific events, but rather a feature that speakers can mobilize for communicative purposes. Finally, we prove how speech events of the same type can end up displaying different interactional qualities as a consequence of the methodological protocol employed in data collection. In general, we argue that fitting speech events into aprioristically determined definitions of spontaneity is often done at the expense of the recognition of speakers' agency.

3.1. Spontaneity as a structural feature

When spontaneity is defined in structural terms, it is not clear which feature (if any) is necessary and sufficient to uniquely identify speech as spontaneous. One could try to derive a list of these features by contrast with lab-produced speech, but phoneticians point out that, even in controlled experimental environments, speakers rarely exhibit features traditionally associated with contrived data (Xu 2010). Hence, features of contrivance are not enough to identify elicited speech events, but the definition could still maintain discriminating power in the opposite direction, that is, it could be that their presence is enough to qualify a speech event as non-spontaneous. To check whether this is the case, let us consider the conversational behavior of two of the most widely-recognized (Xu 2010) hallmarks of non-spontaneous speech: 1) slow speech rate and 2) hyperarticulation.

Isolated occurrences of hyperarticulation and slow speech rate are used in conversation to signal errors and repairs (Biro *et al.* 2022). This function can be maintained in extended instances, in specific cases like the one presented in Text 3. In this excerpt, Luca (native in Italian) is visiting his grandmother Pina. Pina employs hyperarticulation to repeat and correct the words her grandson mispronounces in Bustocco. Luca employs hyperarticulation to accommodate his grandmother, who has suffered from hearing loss, by repeating portions of speech that she has not heard. In Text 3, Luca asks Pina about the last time she visited a mountain peak she used to visit with her late husband.

TEXT 3

1	PINA;	*E qui*	And here
2		*la pazienza*	(we need) patience
3		*sem qui [inscì]*	we are here like this
4	LUCA;	*[L'ultima volta] che te s'è andà su al Tornion qua-*	the last time you went up the Tornion wh-
5		*quando l'è stata*	When was it
6		*(0.2)*	
7	PINA	*eh*	uh
8	LUCA;	*%L'ultima volta*	The last time
9		*Che te s'è andà su*	That you went up
10		*Al Tornion%*	The Tornion
11	PINA;	*Ah il nonno è morto nel dieci*	Ah your grandfather died in (twenty-)ten

As Pina is recounting her impossibility to go hiking because of health issues (IUs 1–3 are the ending portion of the turn), Luca asks about the last time she has been on a specific mountain peak (IUs 4–5). Pina does not understand the question and asks for repetition (IU 7). At this point, Luca raises his voice, slows down the speech rate, and hyperarticulates the elements he is producing. He splits the original content of IU 4 (*the last time you went up the Tornion*) into three IUs (8–10), each one containing a syntactic constituent (*the last time*, *that you went up*, *to the Tornion*), which he hyperarticulates. This is done to facilitate Pina's comprehension and, when she provides an answer (IU 11), the exchange resumes.

Text 3 features an unplanned informal speech event that routinely takes place in the participants' lives, and yet, it extensively presents hallmark features of non-spontaneous speech. Though there are cases in which these features are in fact encountered in non-spontaneous speech, they can also occur as a consequence of accommodating to the communicative conditions of the interaction. This strategy is used by adults interacting with children (Kuhl *et al.* 1997; Uther *et al.* 2007), caregivers with elders (Kemper 1994), and native speakers with L2 speakers (Kangatharan *et al.* 2021). In other words, employing hyperarticulation and slow speech rate is a productive interactional strategy adopted in situations with an unbalance of linguistic competence. Eliminating these linguistic features from the corpus may lead to a reduced visibility of the speakers who produce them, a limitation which corpus compilers are aware of and deal with in different ways. Ultimately, if we deem linguistic accommodation to be a usual occurrence in the life of many language users, then it is of value to include it in a corpus of spoken

discourse. This was a particularly pressing issue in the case of the MULTCORSKL because of the linguistic landscape in which we are operating. Because Kazakhstan displays a large-scale institutionalized multilingualism with age stratification (Agbo and Pak 2017), cases of linguistic accommodation are a frequent occurrence of life. This is by no means unique to Kazakhstan, yet representation of L2 speakers tend to be confined to specialized corpora.

3.2. Spontaneity as a genre feature

An alternative definition of spontaneity conceptualizes it in terms of lack of planning of an event. In this approach, spontaneity is evaluated along a continuum and treated as a feature of the genre (or register) of a specific speech event (Swales 1990; Blackwell and White 2018). Social activities are arranged in relation to each other according to the level of spontaneity that is allowed by the norms regulating them. These norms maintain a certain degree of stability within a culture and can vary cross-culturally. For example, religious rituals in the Catholic world are rigorously planned both in the structural order of the sequences to be performed and in the words to be used (Szuchewycz 1994). Shamanic rituals in Buryatia (or elsewhere in Russian Siberia) may include relatively unplanned exchanges between the petitioner and their ancestor' spirits (Quijada *et al.* 2015; Nagy 2016).

In addition to cross-cultural differences, the association between degree of planning of an individual speech event and genre is not guaranteed even within the same cultural context. Within the same genre, different traditions may have an impact over the shaping of the single event. For example, a conference presentation for a historian entails the reading of a written document prepared in advance, a practice that, at most linguistics conferences, would signal someone as either a novice or an outsider. Even variations at the individuals' level can have an impact over the degree of planning of an event. Recent years saw a rise in social media content about mental health (Haq *et al.* 2022) that led terms and tools from psychology to sip into the everyday vocabulary of certain demographics (Scherlis 2023). Among these tools, there is a script for conflict management called 'I-statements' (Rogers *et al.* 2018), for instance*, I felt ignored* instead of *you ignored me*. The popularization of this communicative script makes it so that, even within the same demographic, an instance of the 'couple fight' can either play out as a potentially volatile improvised event or as a carefully planned exchange featuring pre-

ordered sequences of grammatical structures that have been selected in advance.

Furthermore, degree of planning can vary even within the course of the same speech event, as there is no guarantee that what started as an event of a specific genre will continue as such (Schegloff 1988). This is especially visible inside conversation, which functions as an interactional matrix where participants can embed texts of different nature. For example, in Text 4, extracted from the SBCSAE, two women are discussing the impeachment of former US president Bill Clinton.

TEXT 4

216	MAUDE;	Okay,
217		Here it is,
218	LONI;	[Oh].
219	MAUDE;	[I] got it a little backwards.
220		(...) (H) (%)
221		Uh:,
222		This is Article Two,
223		Section Four.
224		<READ (...) (H) The President Vice-President and all civil officers of the United States,
225		shall be removed from office on impeachment for,
226		(H) a:nd conviction o:f,
227		(.) (%) treason,
228		(.) bribery,
229		(H) or,
230		(.) other high crimes,
231		(...) and misdemeanors. READ>

In the midst of the conversational exchange, Maude retrieves a law handbook and moves from a stretch of unplanned speech (IUs 216–223) into reading the definition of *impeachment* to her interlocutor (IU 224 onwards). The written excerpt was planned and produced way before the moment of the interaction, and it is used by a participant to construct an argument in favor of her position. In this case, we can observe a movement from unplanned to planned discourse. A movement in the opposite direction can be seen in Text 5, where the speaker inserts a chunk of unplanned speech into an otherwise planned event.

TEXT 5

1	GINETTO;	*<RECITE tuta sta beleza,*	all this beauty,
2		*la ma muisna ul coeui,*	makes my heart tender,
3		*E men,*	and I,
4		*Ca iu soi da buen ora RECITE>,*	who have been here for an hour,
5		*(.)*	(.)
6		*<L2 C'è un errore qui scusa >?*	Excuse me there is a mistake here?
7		*(.)*	(.)
8		*Ma muisna ul coeui?*	Makes my heart tender?
9		*/a mən/.*	/a mən/.
10		*<L2 io ho detto >,*	I said,
11		*/e mən/.*	/e mən/.
12		*<L2 Possiamo [interro][₂mpere:],*	Can we stop,
13	GIORGIA;	*[mhm],*	mhm
14		*[₂si certo].*	sure.
15	GINETTO;	*ripartire:,*	restart,
16		*C'- c'è [₃una pa]usa,*	there is a pause,
17	GIORGIA;	*[₃assolutamente]>.*	absolutely.

In this case, the speaker, Ginetto Grilli, is a renowned poet writing in Bustocco. During what the first author originally intended as a recording session of a conversation, Ginetto offers instead to recite a series of poems. During the declamation, he moves between planned recital speech and conversation to comment on language issues. In this excerpt, we can observe one instance where he interrupts the recital to repair his performance (IU 5). Same as Maude, Ginetto is merging material with different levels of planning for communicative purposes.

A potential objection to our criticisms to the unreliability of the concept of spontaneity is that corpus designers can always suspend the application of structural definitions in cases like the exchange between Pina and Luca (Text 3) and ignore local instances of planned speech in the case of Maude (Text 4). We do not doubt that this is true, but we contend that the issue with a structurally-based or a genre-static definition of spontaneity lies in the number of 'exceptional' cases that one needs to account for, in the nature of those 'exceptional' cases (Hall 2008), and in what they reveal about which speakers and genres get to set the standard of what counts as linguistic data of interest. Moreover, we contend that definitions which leave up to the individual researcher the interpretation of (frequent) deviations from the norm require more work to explain the outliers than they provide heuristics for the selection of spontaneous speech.

3.3. Spontaneity and data collection protocols

The final issue with the concept of spontaneity lies in the fact that it can be impacted by the methodology employed for data collection, which makes it difficult to compare events of the same type across corpora. For example, consider two instances of spontaneous phone conversations. Text 6 is taken from the *Switchboard Corpus* (Godfrey *et al.* 1992). In this corpus, participants chose some topics of conversation from a list of prompts and were matched by a robot operator with a stranger that had selected the same interests.

TEXT 6

Topic 303:
THE TOPIC IS CLOTHING. PLEASE FIND OUT HOW THE OTHER CALLER TYPICALLY DRESSES FOR WORK. HOW MUCH VARIATION IS THERE FROM DAY TO DAY? HOW MUCH VARIATION IS THERE FROM SEASON TO SEASON?
B: okay hi
A: hi um yeah I'd like to talk about how you dress for work and and um what do you normally what type of outfit do you normally have to wear
B: well i work in uh corporate control so we have to dress kind of nice so i usually wear skirts and sweaters in the winter time slacks i guess and in the summer just dresses
A: um-hum
B: we can't even well we're not even really supposed to wear jeans very often so it really doesn't vary that much from season to season since the office is kind of you know always the same temperature
A: and is right right is there is there um any is there a like a code of dress where you work do they ask

In Text 6, both interlocutors display typical features of unplanned informal interaction such as disfluencies (*um*), floor holders (*well*, *you know*), and contractions (*doesn't*, *I'd like*, *can't*). In terms of conversational organization, backchanneling is present (*um-hum*), but turns are rather long, and overlapping is limited. Similarly to the non-naturally occurring conversation in Table 1, participants coalesce around fixed interactional roles: A takes on the role of questioner and B that of responder. This strategy gives participants a structure to maintain engagement. Compare this dynamic with the exchange in Text 7, extracted from the SBCSAE, which features spontaneous naturally occurring discourse.

TEXT 7

1	>ENV:	... ((RING)) ... ((RING))
2	JILL:	... Hello=,
3	JEFF:	.. How's my favorite girl in the world.
4	JILL:	(H) Hey .. ba=by.
5	JEFF:	.. Who's —
6		(H) Who's the girl that .. I love?
7	JILL:	@@@@[@]

8	JEFF:	[Who's] the girl that I'll do anything for?
9	JILL:	@@@@ .. [$_2$(H)]
10	JEFF:	[$_2$I'll] wash her feet with my [$_3$mouth].
11	JILL:	[$_3$@@]@ @Aw= @honey,
12		[$_4$(H) how] are you[$_5$=].
13	JEFF:	[$_4$@@@]
14		[$_5$after] a ten mile run.
15	JILL:	Ew.
16		.. [@@]
17	JEFF:	[I would].
18	JILL:	.. Ew.
19	JEFF:	I would.
20	JILL:	Ew.

In Text 7, turns are considerably shorter than in Text 6. Overlap between participants is common and engagement is supported not by interactional roles, but by dialogic resonance, the repetition of linguistic structures across different intonation units (Du Bois 2014). Here, Jeff and Jill have similar hierarchical standing in the exchange, and this results in an event with different interactional features than Text 6.

The excerpts in Text 6 and Text 7 are both spontaneous but sourced through different research protocols. Consequently, participants display different levels of engagement in the event. Participants in the *Switchboard Corpus* excerpt have been exposed to structures and lexical choices in the written prompt, which may have primed them to use these structures. As it was for Text 2, participants will likely not engage with the opinions of their interlocutor beyond the current exchange, which results in them falling back onto rigid interactional structures to sustain engagement. Lack of engagement from participants has also been cited as the reason why the *Switchboard Corpus* displays different dynamics of information flow when compared to other corpora of English (Wasow 2002). Despite these considerations, data from the *Switchboard Corpus* are still routinely employed in studies of spontaneous discourse in a wide variety of disciplines because it is considered that they are "natural (appeared in spontaneous speech), can form an intonational phrase, and [their] frequency [is] not driven by a specific and uncommon genre" (Arnon and Snider 2010: 71). The sustained popularity of the *Switchboard Corpus* suggests that corpus users are keen on using conversational data and that there is indeed a need for more data of this kind. However, conversational data should not be sourced at the expenses of interactional dynamics.

4. RECORDING NATURALLY OCCURRING DISCOURSE

Instead of focusing on linguistic structures, one may want to place their focus on the users of a language instead. Speech events do not exist outside of speakers' practices (Duranti and Goodwin 1992) and, as such, they do not exist outside of speakers' lives. We propose that representation of participants' lives can be achieved by building a corpus of naturally occurring discourse. For a speech event to be considered 'naturally occurring' it must happen for the motivations of the participants, have consequences on their lives beyond the moment of the recording, and take place even if it was not going to be recorded (Du Bois 2003). Adopting this framework, only recordings that are culturally, socially, and interactionally relevant to the participants can be included in the corpus.

The 'cast the net wide' model was first successfully implemented by the SBCSAE (Du Bois *et al.* 2000). In this section, we discuss the ways in which the team of the MULTCORSKL adjusted its own original protocol to ensure that the corpus would contain only naturally occurring discourse. We mainly had to reconsider: 1) the level of agency that we wanted to grant participants, 2) the role of researchers in the community and in the events, and 3) the place of conversation in the corpus. Other adjustments related to the use of remote data collection protocols due to the impact of COVID-19, as well as smaller adjustments that are specific to the Kazakhstani geography, will not be discussed here (Troiani *et al.* 2022). We structured this discussion around the points that required a reconsideration of our theoretical assumptions about the nature of discourse, because we think they can offer elements of reflection to teams setting out to build a corpus.

4.1. Discourse and the agency of participants

The largest innovation introduced by the 'cast the net wide' framework is the notion that the only criterion for inclusion of speech events in a corpus should lay in the participants' motivations. The reason for this choice has been explained in Section 1, namely, speech events that are inconsequential to participants' lives yield artificial interactional dynamics that, in turn, have effects on grammar. In the multilingual context of Kazakhstan, one of the immediate ways in which lack of motivation became visible in grammar was the erasure from the data of the otherwise daily occurrences of code-switching with Russian, Mandarin Chinese, Uzbek, and English. Participants tended to control the amount of code-switching they produced and explicitly ascribed the reason to the fact that the team recruited them to participate in a corpus of Kazakh language. Assuming that the research

team was after exclusively 'pure Kazakh', they restrained themselves from producing code-switching. In response to this, we began introducing the project to potential participants as an endeavor to capture instances of life in Kazakhstan, rather than instances of Kazakh language. Far from being trivial, this shift resulted in the preservation of everyday multilingualism in the speech events.

A somewhat technically demanding adjustment was the decision to relinquish control over the speech event to the participants themselves. Wherever possible, we trained the participants to handle cameras and recorders and left it up to the individual to record events happening in their life. This protocol had been successfully employed in the SBCSAE. The implementation of such a protocol requires researchers to spend time training participants, to engage in consistent communication with participants to ensure the smooth operation of the recordings, and an availability of time (to train participants in the handling of equipment) and material resources (to distribute equipment across participants). When delegating the recording in its entirety was not possible, we let participants propose to us the events they wanted to have recorded. In our case, enlisting the help of participants in the planning and recording of events was worth the effort, especially in cases of cultural significance, culture-specific restrictions, or religious value that the presence of researchers would have disrupted, and which could easily be attended by members of the community that usually participate in these events.

Perhaps the most complex adjustment to be implemented in order to ensure the full agency of participants was accepting that they could bring and impose their own expectations of the event in the recording session. For example, consider the following interaction, which features the first author and four speakers of Bustocco, all women above 60. The first author had originally intended to collect some sociolinguistic information about the participants, set up the recorder, and leave them alone. Consent had been collected days prior to this event. After one of the participants mentions that she was acquainted with a relative of the first author, the whole event is re-casted by the participants as an interview, as can be seen in Texts 8 and 9.

TEXT 8

1	CINZIA;	*io ho vent'anni in piu'?*	I am 20 years older?
2		*Hai visto [la mia data],*	you saw my date (of birth),
3	GIORGIA;	@@	@@
4	BARBARA;	*[oh ascolta],*	hey listen,
	> CINZIA		
5		*Lascia parlare –*	let (-) speak –

6		*Com'è che si chiama.*	What's your name.
7	GIORGIA;	*Giorgia.*	Giorgia.
8	BARBARA;	*Eh lascia parlar la Giorgia adesso.*	Let (the) Giorgia speak now.
9	GIORGIA;	*@@@*	@@@
10		*No,*	No,
11	CINZIA;	*eh.*	what.
12	BARBARA;	*lascia parlar [la Giorgia adesso].*	Let (the) Giorgia speak now.
13	GIORGIA;	*[va benissimo],*	It's fine,
14	BARBARA;	*e' lei che deve farti le domande,*	she is the one that has to ask you the questions,
15	ANNA;	*se deve dirci qualcosa ce lo dica?*	If you have to ask something let us know?

In Text 8, as Cinzia is in the middle of a complaint (IUs 1–2), Barbara interrupts her to demand that the first author be allowed to speak (IUs 4–6). Barbara repeats the request a few times (IUs 5, 8, 12), while the first author overtly states her intention not to be part of the interaction (IUs 9–10 and 13). At this point, Barbara explicitly details her expectations about the conversational roles in the event, and namely that the first author is the person in charge of asking questions (IU 14). Anna affiliates with Barbara by assuring the first author that she is in charge of directing the interview as she sees fit (IU 15). Barbara, Anna, and Cinzia frame the event as an ethnographic interview, and they hold each other accountable for their role as interviewees in the exchange. The research tools (sociolinguistic questionnaire) that the first author brought into the event are also actively re-purposed to fit the participants' interactional goals. For example, consider Text 9.

TEXT 9

BARBARA;	*C'è anche un cimitero che è stato- –*	There was also a cemetery that has been –
ANNA;	*ecco e allora lì c'era una casetta,*	Right and then there was there a little house,
CINZIA;	*cià andiamo avanti,*	Come on let's move on,
> GIORGIA		
	<L2 ndem innanzi >,	<L2 Let's move on >,
ANNA;	*c'era qualcosa?*	Was there something?
	Lì dove han fatto –	There where they did –
CINZIA;	*<READ usa il bustocco >,*	< READ do you use Bustocco>,
GIORGIA;	*quanto spesso usa il bustocco.*	How often do you use Bustocco.
CINZIA;	*spesso spesso,*	Often often,
	Spessissimo,	Very often,
GIORGIA;	*tutti i giorni?*	everyday?

As other participants are still engaging the previous question, Cinzia grows impatient,

leans in on the form that the first author is holding in hand, and requests her to continue with the next question. Cinzia reads the question herself and provides a short response that she will later expand into a narrative. At this point in the exchange, the sociolinguistic questionnaire has become a prop for the interlocutors to continue their conversation and the first author adjusted to meet the participants' expectations by asking questions when provided with the opportunity.

The event in Texts 8 and 9 would have not been recorded, had the first author not given up on her expectations about how a conversation between friends should look like. After the recorder is set up and one of the participants instructed to push the button wherever comfortable, as the first author is getting ready to leave, Cinzia explicitly motions her to sit and fill out the questionnaire. The first author's positionality leads the participants to assume that her interactional role in the exchange is that of interviewer. Specifically, the fact that a participant is acquainted with the grandmother of the first author contributes to the re-casting of the event in terms that are familiar with the participants.[7] As Cinzia and the first author move through the form, they adhere to the 'school interview' expectation that Cinzia has in mind. In response to the questions, Cinzia offers long narratives that the other participants end up co-constructing through commentaries, addition of details, banter, and comparisons to their own life experience. About 20 minutes into the event, the first author finally accepts that the event that is unfolding is indeed an instance of naturally occurring conversation and records the following 3.5 hours. The data that resulted from this session is a conversation among four friends embedded into an overarching sociolinguistic interview. The sociolinguistic interview is not by itself a genre that happens for the sake of participants' motivations, but this speech event shows how participants and researchers can co-construct the naturalness of an event by letting participants' agency take over the event. In this case, participants engaged in the interaction with the goal of telling their life stories to a researcher, member of the community, who could have easily been one of their grandchildren. Adjusting to these expectations resulted in the collection of a naturally occurring event.

---

[7] Participants refer to the doctoral program in which the first author is enrolled at the time of the research as 'school', and they mention having already participated in assignments where students are asked to interview older members of their family.

4.2. Discourse and the positionality of the researcher

Adopting the motivations of participants as base criterion for data collection also requires reconsidering the figure of the researcher and the role of the researcher in the speech event. Scientists are traditionally encouraged not to participate in recordings to preserve data authenticity (Potter 2002). This observation makes sense in a context where the researcher is not a member of the community, because the interactional dynamics of the exchange are artificially constructed. The consequences of merging the roles of community members, researchers, and participants can be particularly well appreciated in elicited data. Consider for example the following two excerpts featuring two varieties of Mixtec. In Text 10, Jeremías Salazar is retelling a childhood memory to the first author. The narrative is in Sà'án Sàvǐ ñà Yukúnanǐ, a language the first author does not speak nor understand.[8]

TEXT 10

1	JEREMIAS;	*kuě níkùu tavǎ-kue-yì-tí cha ñàà*	they could not take it out
2		*Tsà'ǎ ñà-ka*	the reason is
3		*tí luu-ni kìsi luu-ka ra nìkee viǐ-ni chùún-ka cha tá nchò'ǒ-tí ra*	the pot was so little that the chicken barely fit so when it cooked
4		*ñàà kuě níkùu tavǎ-kue-yì-tí cha saán kúu-ñà ñàà*	they could not take it out that is when
5		*Ntsà'àn-kue-yì ra ñàà*	they went and
6		*nìntà... nì*	he
7		*nìkǎni-à mátsá'nù tavà ñàà*	he called my grandma so (then)
8		*nchìnchiǐ-kue-yì nixi sǎ'a-kue-yì ñàà távî chùún-ka*	she would help them get the chicken out
9		*cha tíí saán ntsàǐ ra ñàà*	probably she just got there
10		*ntsǎ'nchǐ chùún-ka tíí ñàà*	she cut the chicken

This narrative was elicited in the context of a documentation project. Jeremías chose the topic in advance, but he did not script the content, which resulted in a somewhat unplanned speech. This excerpt features disfluencies (Belmar and Salazar 2023) and repairs (IU 6), but no backchannel from the first author. Jeremías does not attempt to include the first author in the exchange. The elicitation protocol presented here is not uncommon among documentary linguists and it yields a considerably different result when the event is elicited by a community member. In Text 11, Juan Miranda is telling a narrative in Tù'un Na Ñuu Sá Matxí Ntxè'è to a young woman member of his community

---

[8] Materials are accessible at https://sites.google.com/view/saansavi-yucunani (accessed 25 September 2023).

and speaker of the variety.[9]

TEXT 11

1	JUAN;	*xa kasantxeena xa sa'ana lucha ñi'i ñi'itana shu'un*	they are already sowing they are already struggling they are finding finding money
2		*ta chixin madre nte'i chi koo mi ña koo ña'a vi*	and before there were no mothers who were poor there are no well
3		*unkivi koo mi modo kakio*	we cannot save ourselves in any way
4		*ta vitxi*	and now
5		*ta kachi ji'un*	as I told you
6		*Ntxe'e*	look
7		*yo'o*	you
8		*Kunte'ivo ñaa*	we(.incl) are poor because of this

As in the case with the first author in Text 10, Juan's interlocutor never claims the floor and provides only occasional backchannel. But in this case, she is a plausible audience for the event and the lack of overt verbal engagement does not prevent Juan from involving her in the event through various strategies. In IU 5, he recalls a previous point in the conversation to establish that he and the interlocutor share joint knowledge (*as I told you*). In IU 6, he uses an attention-grabbing expression (*look*) and in IU 7 he overtly addresses the interlocutor indicating her as a recipient of the narrative through the use of a second person pronoun (*you*). Finally, in IU 8 Juan shifts to an inclusive first-person plural (*-o* in *kunte'iv-o*) and recruits the interlocutor in the narrative (*this is why we are poor*) acknowledging that the interlocutor has experience on the topic of the narrative, namely, the economic conditions of the Pueblo of San Martin Duraznos.

The narratives in Texts 10 and 11 have both been elicited, but Text 11 has been elicited enlisting community members as recorders and creating a condition where participation in the event would be plausible. This gives Juan a motivation to ensure the interlocutor understands the story, which results in a significant alteration of grammatical structures employed in the event. The data resulting from the elicitation in Text 11 are not 'naturally occurring', yet they can be considered naturalistic, that is, elicited data that attempt at recreating the interactional dynamics that would have been at play in naturally occurring discourse (e.g., by ensuring that data are elicited by a member of the speaker's community). While accessing naturally occurring data may not always be possible (e.g.,

---

[9] The original transcription was translated into Spanish. The English translation was done by the first author. Interactional elements, disfluencies, and truncation are not transcribed in the original data.

in communities with a small number of speakers), a valid alternative is to employ elicitation protocols which make an explicit point about maintaining interactional dynamics. One such model is the Pear Story elicitation protocol in its original formulation (Chafe 1980) or in the modified version featuring a three-party interaction (Kibrik and Fedorova 2018).

If the recorder is recognized as a member of the community, their refusal to intervene could alter the dynamics of the event more than letting themselves be recruited into the participants' plans. This recruitment can play out in extreme forms like the case of the first author being involved in the entire event of Texts 8 and 9 are taken, or it can materialize in shorter occurrences like the one in Text 12. Here, the researcher (Serik) is operating the camera and recording a conversation between his grandparents. The event unfolded for a couple of hours. Aisha and Adilet have been conversing in a room by themselves, Serik occasionally checks on the camera.

TEXT 12

1	AISHA;	*Taksïmen mına jazwım bitse qaytıp ketemiz dep otır*	He is saying he will go back to the city by taxi as soon as he is done with his recordings
2		*Qalağa (H)*	To the city
3		*endi voobşçe ketemin deydi ğoy*	He is saying he will be gone for good
4		*Sen oqwğa ketesiŋ be*	Are you leaving for school?
5		*Astanağa*	To Astana
6	ADILET;	*Oqwın bitte emes pe ey*	Don't you know he has already finished his studies
7	SERIK;	*Altısında ketemin*	I'll leave on the sixth
8	AISHA;	*A.*	Come again?
9	SERIK;	*Altısında ketemin*	I'll leave on the sixth
10	AISHA;	*Altısında*	On the sixth?
11	ADILET;	*Astanağa*	To Astana
12	AISHA;	*Ïä ket- ketedi Astanğa*	Yes, he will leave for Astana

This excerpt comes after a period of silence in which Aisha is knitting and Adilet is resting. Upon noticing Serik close to the camera, Aisha tells Adilet that their grandson is about to return to leave town (IUs 1–3). Aisha addresses Serik directly and asks whether he is returning to school (IUs 4–5). Serik responds (IU 7 and 9) and Aisha and Adilet resume their conversation. In this case, Serik's involvement in the recording is not only plausible, but also necessary to maintain a natural dynamic. Serik is the researcher, but he is also Aisha's grandson, and it is in this capacity that his intervention is required. Participation of researchers into recordings is a practice that was not actively encouraged

within the MULTICORSKL, but it was not prohibited when the researcher's presence in the event is usual.

4.3. Discourse and conversation

The final adjustment needed to obtain a corpus of naturally occurring discourse was the reconsideration of the role of conversation within the corpus. The importance of conversation cannot be understated. Conversation is the most basic form of communication available to people and a universal of human language (Schegloff 2015). It is the vehicle through which speakers acquire, maintain, and modify grammar (Thompson *et al.* 1996; Lytle and Kuhl 2017). Moreover, the status of conversation as a standalone genre is contested by many scholars (see Warren 2006 for an overview of the debate). As seen in Section 3.2, conversation is best understood as an interactional matrix within which other genres are embedded as needed. For these reasons, conversation is a non-negotiable component of our corpus, a position that has been explicitly supported by other corpus designers, whether with the intention to cater to users' needs (Love *et al.* 2017: 324), highlight the role of spoken discourse in language acquisition (Raso and Mello 2014), or maximize resource efficiency by selecting data that are as different as possible from written discourse (Čermák 2009).

While the fundamental role of conversation in human life should not be understated, a corpus representing life cannot limit itself to conversation alone. Corpora like the SBCSAE and the MULTICORSKL record language in use for the motivations of the participants and source speech events that are consequential to the lives of participants. Though grammar is acquired in conversation, many life goals are achieved by genres other than conversation or otherwise spontaneous discourse. This includes goals that are not concretely measurable, such as the competent performance of culture and identity (Dingemanse and Floyd 2014). For this reason, the MULTICORSKL does not limit itself to conversation and includes a series of recordings featuring poem recitals, ritualistic animal slaughters, political speeches, interviews to culturally relevant figures, guided visits to historical sites, etc. We hope in this way to gain a representation of language in use that is as faithful as possible to the lives of speakers of Kazakh.

5. CONCLUSIONS

In this paper, we presented the theoretical assumptions underlying the 'cast the net wide' framework, an approach to data collection that aims at the compiling of corpora of naturally occurring spoken discourse. This framework posits as the goal for a corpus to be representative of the life of participants rather than of the linguistic structures in use. The reason for this decision is demonstrated in Section 1, where we present two case studies (backchannel and question-answer sequences) showing how the frequency and distribution of interactional features, as well as the structuring of interaction, varies as a result of whether the speech event under analysis is naturally or non-naturally occurring. The result of this analysis suggests that differences between speech events go beyond their genre and presence or lack of spontaneous speech.

We turn to demonstrate the issues with relying on the notion of spontaneity in Section 3. First, we show that the structurally based definition of spontaneity is not a reliable heuristic for the detection of unplanned speech, in particular when considering interactions that feature linguistic accommodation in situations of power unbalance (e.g., native speakers talking to L2 speakers, caregivers to elderly or children). Then we demonstrate how spontaneity is not a fixed quality of a genre, as language users can mix genres and degrees of planning to achieve their goals. The data presented suggest that spontaneous conversation overall acts as an interactional matrix where interlocutors can embed less spontaneous genres for their own reasons. Finally, we demonstrate how the same speech event can result in an exchange of varying spontaneity as a consequence of the research protocol employed in recording it. All these observations motivated us to define an alternative criterion for the selection of data in a corpus of spoken discourse.

As an alternative to the selection of structural criteria such as spontaneity, we propose an approach to corpus design that prioritizes the collection of naturally occurring discourse. We define naturally occurring speech events as events that happen for the social and interactional goals of the participants and would take place even if there were not going to be recorded, and, as such, are consequential to the participants' lives beyond the moment of the recording. We suggest that there are a few adjustments corpus designers can introduce to their workflow if they are interested in sourcing naturally occurring discourse. In particular, we suggest that participants are recruited as recorders and are given the opportunity to bring their understanding of the event. We also suggest that researchers who are present to an event comply with the understanding of the interactional

dynamics displayed by participants and involve community members in the data recording phase. We argue that these adjustments can help researchers approximate, as best as possible, the representation of language as it occurs in the lives of its speakers.

REFERENCES

Agbo, Seth A. and Natalya Pak. 2017. Globalization and educational reform in kazakhstan: English as the language of instruction in graduate programs. *International Journal of Educational Reform* 26/1: 14–43.

Arnon, Inbal and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62/1: 67–82.

Auderset, Sandra and Carmen Hernández Martínez. 2021. Documenting Tù'un Na Ñuu Sá Matxí Ntxè'è, a mixtec language of Oaxaca, Mexico. *Endangered Languages Archive*. http://hdl.handle.net/2196/a3085a77-687a-48b9-9caf-a48c3c1f1f1f.

Biro, Tifani, Annie J. Olmstead and Navin Viswanathan. 2022. Talker adjustment to perceived communication errors. *Speech Communication* 138: 13–25.

Blackwell, James W. and Peter R. R. White. 2018. The building blocks of speech: Spontaneity, pre-packaging and the genre structuring of university lectures. *Text & Talk* 38/3: 267–290.

Burnard, Lou. 2002. Where did we go wrong? A retrospective look at the British National Corpus. In Bernhard Ketteman and Georg Marko eds. *Teaching and Learning by Doing Corpus.* Amsterdam: Rodopi, 51–70.

Čermák, František. 2009. Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics* 14/1: 113–123.

Chafe, Wallace L. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Westport: Praeger.

Chui, Kawai and Huei-ling Lai. 2008. The NCCU corpus of spoken Chinese: Mandarin, Hakka, and southern Min. *Taiwan Journal of Linguistics* 6/2: 119–144.

De Fina, Anna and Sabina Perrino. 2011. Introduction: Interviews vs. 'natural' contexts: A false dilemma. *Language in Society* 40/1: 1–11.

Dingemanse, Mark and Simeon Floyd. 2014. Conversation across cultures. In N. J. Enfeild, Paul Kockelman and Jack Sidnell eds. *The Cambridge Handbook of Linguistic Anthropology*. Cambridge: Cambridge University Press, 447–480.

Drummond, Kent and Robert Hopper. 1993. Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on Language & Social Interaction* 26 2: 157–177.

Du Bois, John W. 2003. Discourse and grammar. In Michael Tomasello ed. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure.* London: Lawrence Erlbaum Associates, 61–102.

Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25/3: 359–410.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson and Nii Martey. 2000. *Santa Barbara Corpus of Spoken American English*. Philadelphia: Linguistic Data Consortium.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming and Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards and Martin D. Lampert *Data: Transcription and Coding in Discourse Research.* London: Lawrence Erlbaum *Talking*, 45–89.

Du Bois, John W. and Giorgia Troiani. 2022. *Cast the Net Wide: Corpus as a Slice of Life*. (Presentation, 25 February 2022*)*. Bologna: Italy.

Duranti, Alessandro and Charles Goodwin. 1992. *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press Cambridge.

Filchenko Andrey, Giorgia Troiani, John W. Du Bois, Gulnar Sarseke, Akyl Akanov, Moldir Bizhanova, Nikolay Mikhailov, Tansulu Temirbekova, Bybaris Seitak and Zhansaya Turaliyeva. 2023. *Multimedia Corpus of Spoken Kazakh Language* (version 1).

Godfrey, John J., Edward C. Holliman and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* San Francisco: IEEE Computer Society, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Greenbaum, Sidney. 1991. The development of the International Corpus of English. In Karin Aijmer and Bengt Altenberg eds. *English Corpus Linguistics: Studies in Honour Svartvik*. London: Longman, 83–91.

Hall, Kira. 2008. Exceptional speakers: Contested and problematized gender identities. In Janet Holmes and Miriam Meyerhoff eds. *The Handbook of Language and Gender*. New York: Wiley Blackwell, 353–371.

Haq, Ehsan-Ul, Lik-Hang Lee, Gareth Tyson, Reza Hadi Mogavi, Tristan Braud and Pan Hui. 2022. Exploring mental health communications among Instagram coaches. In Nitin Agarwal, Zongmin Ma and Jon Rokne eds. *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* New York: IEEE Press, 218–225.

Heinz, Bettina. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 357: 1113–1142.

Hernández Martínez, Carmen, Griselda Reyes Basurto and Eric W. Campbell. 2021. MILPA (Mexican Indigenous Language Promotion and Advocacy): A Community-centered linguistic collaboration supporting indigenous Mexican languages in California. In Justyna Olko and Julia Sallabank eds. *Revitalizing Endangered Languages: A Practical Guide.* Cambridge: Cambridge University Press, 216–217.

Kangatharan, Jayanthiny, Maria Uther and Fernand Gobet. 2021. The effect of hyperarticulation on speech comprehension under adverse listening conditions. *Psychological Research* 86: 1–12.

Kemper, Susan. 1994. Elderspeak: Speech accommodations to older adults. *Aging, Neuropsychology, and Cognition* 1/1: 17–28.

Kibrik, Andrej A. and Olga V. Fedorova. 2018. An empirical study of multichannel communication: Russian pear chats and stories. *Psychology. Journal of the Higher School of Economics* 15/2: 191–200.

Kucera, Karel. 2002. The Czech National Corpus: Principles, design, and results. *Literary and Linguistic Computing* 17/2: 245–257.

Kuhl, Patricia K., Jean E. Andruski, Inna A. Chistovich, Ludmilla A. Chistovich, Elena V. Kozhevnikova, Viktoria L. Ryskina, Elvira I. Stolyarova, Ulla Sundberg and Francisco Lacerda. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science* 277 (5326): 684–686.

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

Lytle, Sarah Roseberry and Patricia K. Kuhl. 2017. Social interaction and language acquisition: Toward a neurobiological view. In Eva M. Fernández and Helen Smith

Cairns eds. *The Handbook of Psycholinguistics*. New York: Wiley Blackwell, 615–634.

Nagy, Zoltán. 2016. *The Khanty of Vasyugan. Change of the Religious System in XIX-XXI Centuries*. Tomsk: Tomsk State Pedagogical University Publishing House.

Oostdijk, Nelleke. 2002. The design of the spoken Dutch corpus. In Pam Peters, Peter Collins and Adam Smith. *New Frontiers of Corpus Research*. Amsterdam: Rodopi, 105–112.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling and William Raymond. 2005. The Buckeye Corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45/1: 89–95.

Pomerantz, Anita. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. Maxwell Atkinson and John Heritage eds. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, 57–101.

Potter, Jonathan. 2002. Two kinds of natural. *Discourse Studies* 4/4: 539–542.

Quijada, Justine B., Kathryn E. Graber and Eric Stephen. 2015. Finding 'their own': revitalizing buryat culture through shamanic practices in Ulan-Ude. *Problems of Post-Communism* 62/5: 258–272.

Raso, Tommaso and Heliana Mello. 2012. The C-ORAL-BRASIL I: Reference corpus for informal spoken Brazilian Portuguese. In Vládia Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina Scarton, Fernando Batista, Diego Silva, Catarina Magro and Hugo Pinto eds. *Computational Processing of the Portuguese Language*. New York: Springer 362–367.

Raso, Tommaso and Heliana Mello. 2014. Spoken corpora and linguistics studies: Problems and perspectives. In Raso, Tommaso and Heliana Mello eds. *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins, 1–24.

Rogers, Shane L., Jill Howieson and Casey Neame. 2018. I understand you feel that way, but I feel this way: the benefits of I-language and communicating perspective during conflict. *PeerJ* 6: e4831. https://doi.org/10.7717/peerj.4831.

Salazar, Jeremias, Guillem Belmar, Catherine Scanlon, Giorgia Troiani and Eric W. Campbell. 2021. Bridging diaspora: Technology in the service of the revitalization of Sà'án Sàvǐ ñà Yukúnanǐ. In Eda Derhemi ed. *Endangered Languages and Diaspora.* Berkshire: Foundation for Endangered Languages, 176–185.

Schegloff, Emanuel A. 1988. From interview to confrontation: Observations of the bush/rather encounter. *Research on Language & Social Interaction* 22/1–4: 215–240.

Schegloff, Emanuel A. 2015. Conversational interaction the embodiment of human sociality. In Deborah Tannen, Heidi E. Hamilton and Deborah Schiffrin eds. *The Handbook of Discourse Analysis*. New York: Wiley Blackwell, 346–366.

Scherlis, Lily. 2023. Boundary issues. *Parapraxis.* https://www.parapraxismagazine.com/articles/boundary-issues

Stivers, Tanya, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federicoi Rossano, Jan Peter, Kyung-Eun Yoon and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. In *Proceedings of the National Academy of Sciences* 106/26: 10587–10592. https://doi.org/10.1073/pnas.0903616106.

Stivers, Tanya, Nick J. Enfield and Stephen C. Levinson. 2010. Question-response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics* 42: 2615–2619.

Stivers, Tanya and N.J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics* 42/10: 2620–2626.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge university press.

Szuchewycz, Bohdan. 1994. Evidentiality in ritual discourse: The social construction of religious meaning. *Language in Society* 23/3: 389–410.

Thompson, Sandra A., Emanuel A. Schegloff and Elinor Ochs. 1996. *Interaction and Grammar*. Cambridge: Cambridge University Press.

Tottie, Gunnel. 1991. Conversational style in British and American English: The case of backchannel. In Jan Svartvik, Karin Aijmer and Bengt Altenberg eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 254–271.

Troiani, Giorgia, John W. Du Bois, Gulnar Sarseke, Andrey Filchenko, Ilya Salimzianov, Nikolay Mikhailov, Fatima Moldashova, Akyl Akanov, Moldir Bizhanova, Dameliya Koishybayieva, Aigerim Khamitova, Tomiris Nurgalyieva, Aigerim Seiilbek, Bybaris Seitak, Bota Tursunova and Aruzhan Yelubay. 2022. Remote workflow as educational opportunity: The experience of the Multimodal Corpus of Spoken Kazakh language. *Coyote Papers*: 11–18.

Uther, Maria, Monja A. Knoll and Denis Burnham. 2007. Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication* 49/1: 2–7.

Warren, Martin. 2006. *Features of Naturalness in Conversation*. Amsterdam: John Benjamins.

Wasow, Thomas. 2002. *Postverbal Behavior*. CSLI Stanford: The University of Chicago Press.

Xu, Yi. 2010. In defense of lab speech. *Journal of Phonetics* 38/3: 329–336.

Corresponding author
Giorgia Troiani
Nazarbayev University
Block 7, office 7e.119
53 Kabanbay Batyr
01000
Astana
Kazakhstan
Email: giorgia.troiani@nu.edu.kz