

Introducing the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE)

Elizabeth Hanks, Tony McEnery, Jesse Egbert, Tove Larsson,
Alexander Holmberg, Douglas Biber, Randi Reppen, Paul Baker,
Raffaella Bottini, Vaclav Brezina, Gavin Brookes, Isobelle Clarke

Abstract

This article introduces the first large-scale and freely available corpus of everyday, private American English conversation, the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE). LANA-CASE contains 9,374,787 words (1,330 hours) of conversation across 1,477 texts collected between 2022 and 2024. It includes a total of 1,116 speakers from across the United States. In Hanks et al. (2024), we described the process of corpus compilation. In this article, we focus on documenting the finished product (i.e., the full corpus). We describe the composition of LANA-CASE in terms of the situational characteristics of its texts and the demographic characteristics of its speakers. We also provide guidelines for accessing and using the corpus, including giving an overview of the transcription conventions and metadata files. We hope that this corpus will prove to be a useful resource for research on various aspects of American English conversation as well as a pedagogical tool for classroom activities and assignments.

Keywords: LANA-CASE, spoken corpus, register, conversation, U.S., metadata, transcription conventions

1. Introduction

Conversation is central to everyday life. In fact, all people—across cultures and languages—participate in conversation as part of their daily routine (Nofsinger, 1991). Yet, compared to written language, few corpora exist that aim to represent spoken registers such as conversation. The spoken corpora that do exist have been instrumental in the study of a wide

range of spoken registers and dialects. For example, the spoken component of the British National Corpus (Leech, 1993) and the Spoken British National Corpus 2014 (the Spoken BNC2014; Love et al., 2017) have been invaluable resources for the study of British English conversation (e.g., lexis, grammar, and pragmatics); the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000–2005) has been integral to analyses of spoken American English in a variety of contexts (e.g., book clubs, sermons, lectures, conversations); the Longman Spoken American Corpus (Stern, 2005) has been used in the development of dictionaries and grammars (e.g., Biber et al., 1999); and the spoken component of the Corpus of Contemporary American English (Davies, 2009) has been a useful tool for examining spoken interactions sourced from TV and radio programs.

Despite the usefulness of these existing corpora, none yet exist that exhibit each of the following characteristics: (a) represent everyday, private conversation; (b) represent diverse speakers of American English across the U.S.; (c) are large in size (i.e., contain several millions of words); (d) represent conversational language from the early 21st century; *and* (e) are freely available. Without a corpus exhibiting each of these characteristics, we are unable to systematically study the linguistic features associated with this particular register and variety of English. These characteristics are discussed in terms of their relevance for corpus linguistic research below.

- (a) *A corpus that represents everyday, private conversation* is generally considered to be the most challenging to compile (e.g., Love, 2020, p. 42). This is largely because doing so requires extensive resources such as time to transcribe (even with the support of speech-to-text technology) and funds to remunerate participants, as well as extended consent procedures to enable licensed, ethical access to audio recorded in private contexts (e.g., Burnard, 2002, p. 6). Due to the challenges associated with collecting language from this register, few corpora represent everyday, private conversation. Spoken interactions in movies, YouTube, or TV interviews that involve a greater degree of interlocutor planning are sometimes used as a proxy to examine conversation. However, research shows that everyday, private conversation exhibits features that differ from other spoken registers (e.g., recorded conversation; Larsson et al., 2026). The spoken component of

the British National Corpus (Leech, 1993) and the Spoken BNC2014 (Love et al., 2017) as well as the conversation component of the London–Lund Corpus of Spoken English 1 (LLC-1; Svartvik & Quirk, 1980) and 2 (LLC-2; Pöldvere et al., 2021) are some of the few corpora of everyday, private conversation that allow for the study of this register. Yet, because they were compiled in the U.K., the findings that can be drawn from these corpora are largely specific to British English, which exhibits features that differ from other dialects of English such as American English (e.g., Baker, 2017).

- (b) *A corpus that represents diverse speakers of American English across the U.S.* is particularly challenging to compile because of the large size and population of the country. Corpora of language from YouTube (e.g., the Corpus of North American Spoken English; Coats, 2023) have been used to examine dialectal features across U.S. cities and states, as have corpora that are more specialized to particular sociolinguistic groups (e.g., the Corpus of Regional African American Language; Kendall & Farrington, 2023). The Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000–2005) is one of the few corpora that represent everyday, private conversation (among other spoken registers) across a variety of sociolinguistic groups (e.g., age, region, occupation). Yet, because it is relatively small at about 249,000 words, it is challenging to use quantitative methods to draw generalizable conclusions from this corpus.
- (c) *A corpus that is large in size (i.e., contains several millions of words)* is useful for analyses that seek to generalize their conclusions about the target linguistic feature(s) to a population. The recommended corpus size depends on the research question (Egbert et al., 2022); however, a corpus of several millions of words and several hundreds or thousands of texts may be necessary to achieve robust, stable results for many linguistic inquiries (e.g., Egbert et al., 2025). While some corpora of spoken language meet these considerations for size (e.g., the Corpus of Contemporary American English [Davies, 2009]; the Corpus of North American Spoken English [Coats, 2023]), these corpora do not represent everyday, private conversation. The Spoken BNC2014 (Love et al. 2017) is one corpus of 11.5 million words which does represent everyday, private conversation and it does so on a scale large enough to permit the drawing of generalizations about conversation as a whole. However, as discussed above, the corpus represents British English, which is distinct from American English.

- (d) *A corpus that represents conversational language from the early 21st century* would be timely, particularly given the recent compilation of the Spoken BNC2014 (Love et al., 2017). A corpus representing current conversational language in the U.S. (i.e., conversations collected between 2022 and 2024), following similar compilation techniques, could serve as a point of comparison for the Spoken BNC2014 which represents recent conversational language in the U.K. (i.e., conversations collected between 2012 and 2016). Such a corpus could later be used to compare conversational language from the early 21st century to conversational language in future decades.
- (e) *A corpus that is freely available* would enable public access to a large collection of U.S. language representing everyday, private conversation from the early 21st century for a wide range of applications, such as development of pedagogical materials as well as research on sociolinguistics, lexis, grammar, pragmatics, and register.

In this paper, we introduce the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE), a resource for the study of American English conversation. This corpus is unique in that it exhibits a combination of all five of the characteristics discussed above, which do not exist jointly in any other corpus to date. Namely, LANA-CASE (a) represents everyday, private conversation; (b) represents diverse speakers of American English across the U.S.; (c) is large in size (i.e., contains several millions of words); (d) represents conversational language from 2022-2024; and (e) is freely available.

In compiling the corpus, we followed the framework for corpus design and representativeness proposed by Egbert, Biber, and Gray (2022, pp. 68-121). They define representativeness as “the extent to which a corpus permits accurate generalizations about the quantitative linguistic patterns that are typical in a target language or discourse domain” (Egbert et al., 2022, p. 11). This framework includes describing the target domain (i.e., describing the domain our corpus aims to represent), operationalizing the domain (i.e., establishing feasible corpus boundaries), planning the corpus sample (i.e., planning methods for recruiting participants and collecting data from the U.S. population for inclusion in the corpus), and evaluating the corpus (i.e., evaluating the operational domain in comparison to the target domain, and evaluating the sample according to the operational domain). Through this process, we designed the corpus

with both the representativeness of the register of conversation and the representativeness of speakers of American English in mind.

The first steps in the construction of LANA-CASE are outlined in other publications as an overview of the process of corpus compilation (i.e., describing the target domain, operationalizing the domain, and sampling; see Hanks et al., 2024 for a brief overview and Hanks et al., under review for an in-depth discussion). In the present article, we describe the composition of the final corpus in terms of the situational characteristics of its texts (Section 2.1) and the demographic characteristics of its speakers (Section 2.2). Following this, we provide guidelines for accessing and using the corpus, including an overview of the transcription conventions and metadata about texts and speakers (Section 3). For an evaluation of LANA-CASE—the final step of corpus compilation (Egbert et al., 2022)—readers may refer to Hanks et al. (under review; summarized in Section 2.2).

2. Corpus composition

Everyday, private conversations were collected for inclusion in LANA-CASE by recruiting diverse speakers across the United States. We utilized Public Participation in Scientific Research (PPSR; Shirk et al. 2012), following methods used in the Spoken BNC2014 (Love et al. 2017) and the National Corpus of Contemporary Welsh (Knight et al. 2021). Through this data collection method, participants recorded their own everyday, private conversations and submitted these recordings to the team, along with information about the recording and the accompanying speakers. Participants were not provided prompts about the required length nor topic of their recordings (see Hanks et al., 2024 and Hanks et al., in preparation for more information). Each recording submitted by participants was then transcribed and included in the corpus as a text.

LANA-CASE contains 9,374,787 words¹ (1,330 hours) of conversation stored across 1,477 texts collected between 2022 and 2024. Each text constitutes a transcription of a single audio file that was submitted by participants. LANA-CASE includes a total of 1,116 speakers from across the U.S. The mean speaker contribution is 8,391.7 words ($SD=9,829.4$) across a mean of 2.9 texts ($SD=4.3$, mode=1 text). These data are shown in Table 1 below.

Table 1. Speaker contributions across words and texts

	Words per speaker	Texts per speaker
Mean	8,391.7	2.9
Standard deviation	9,829.4	4.3
Maximum	72,904	85
Minimum	6	1

The corpus features a wide range of situational characteristics (e.g., mode, location, and communicative purpose(s) of the *texts* in the corpus) as well as a wide range of speaker characteristics (e.g., age, race/ethnicity, and geographic region of the *speakers* in the corpus). In order to use the corpus, it is necessary to understand its composition in terms of both of these elements. Below, we present the composition of LANA-CASE according to its situational characteristics (Section 2.1) and speakers (Section 2.2). We also provide brief recommendations for evaluating the corpus according to users’ specific research purposes (Section 2.3).

2.1. Situational characteristics of LANA-CASE texts

Spoken corpora are unique in that their target domain must be specified both in terms of the situational characteristics of the texts and the demographic characteristics of the speakers who

¹ There are many ways to count words, especially in spoken corpora. The numbers we present throughout this article are the most inclusive word counts, which consider merged forms (e.g., “can’t”, “woulda”) as multiple words and include unclear and anonymized speech in the total count. A more conservative word count for the full corpus, which considers a single word as any string of characters separated by white space and does not include unclear nor anonymized speech, is 8,403,376. Both word counts are available in the metadata so users may select which operationalization aligns more closely with their research goals.

produce those texts. In the compilation of LANA-CASE, we collected information about each of these elements. Here, we provide an overview of details related to corpus texts.

Throughout the process of recruiting contributors to the corpus, metadata about the situational characteristics of each recording was collected and noted along with the transcribed corpus text. Details about metadata collection and organization are available in the corpus manual (Authors et al., in preparation). The final composition of LANA-CASE in terms of self-reported situational details is summarized in the tables below. These tables include:

- the number of interlocutors² (Table 2);
- mode (Table 3);
- setting (Table 4);
- interlocutor relationship(s)³ (Table 5); and
- communicative purpose(s)⁴ (Table 6).

Table 2. Number of interlocutors in LANA-CASE texts

Number of interlocutors	Texts	% Texts	Words	% Words
2 speakers	1208	81.8%	7,596,313	81.0%
3 speakers	230	15.6%	1,547,781	16.5%

² Through recruitment, we targeted interactions between relatively few interlocutors in order to facilitate the accurate identification of speakers in the transcription process (see Love, 2020, Chapter 6, pp. 136-163 for information about accurate speaker diarization; see Hanks et al., 2024, p. 27 for information about this decision in relation to the compilation of LANA-CASE). Most texts include 2 or 3 speakers. Some texts that include another number of interlocutors are also available in the corpus, as long as (a) some speakers are not present throughout the entire recording, such as when one speaker enters the recording as another speaker leaves, and/or (b) the speakers are relatively easy to differentiate and exhibit demographic characteristics that had been under-represented in the corpus at the time of submission (see Hanks et al., under review for a discussion).

³ Participants could self-report more than one relationship among interlocutors in the text. Thus, percentages in each row of this table should not be summed. Rather, they should be interpreted independently of one another, where the numerator is the number of texts (or words from these texts) that are tagged for the relationship and the denominator is the total number of texts (or words) in the corpus.

⁴ Participants could self-report more than one communicative purpose of the text. Thus, percentages in each row of this table should not be summed. Rather, they should be interpreted independently of one another, where the numerator is the number of texts (or words from these texts) that are tagged for the communicative purpose and the denominator is the total number of texts (or words) in the corpus.

Other	39	2.6%	230,693	2.5%
Total	1,477	100%	9,374,787	100%

Table 3. Recording mode of LANA-CASE texts

Mode	Texts	% Texts	Words	% Words
Spoken face-to-face	1202	81.4%	7,370,207	78.6%
Spoken remotely	274	18.6%	2,002,148	21.4%
Both face-to-face and remote	1	0.0%	2,432	0.0%
Total	1,477	100%	9,374,787	100%

Table 4. Location of LANA-CASE texts

Location	Texts	% Texts	Words	% Words
Home	703	47.6%	5,009,176	53.4%
Car	201	13.6%	1,058,781	11.3%
Restaurant	46	3.1%	301,229	3.2%
Call	41	2.8%	335,146	3.6%
School	14	1.0%	103,920	1.1%
Office	11	0.7%	105,941	1.1%
Outside	9	0.6%	65,533	0.7%
Public space	12	0.8%	25,256	0.3%
Store	6	0.4%	21,315	0.2%
Any combination of the above	416	28.2%	2,229,092	23.8%
Unknown	18	1.2%	119,398	1.3%
Total	1,477	100%	9,374,787	100%

Table 5. Interlocutor relationship(s) of LANA-CASE texts

Relationship	Texts tagged for this relationship	Texts <i>not</i> tagged for this relationship	% of texts tagged for this relationship	Words in texts tagged for this relationship	Words in texts <i>not</i> tagged for this relationship	% of words in texts tagged for this relationship
Close family/partners/very close friends	1,179	298	79.8%	7,535,509	1,839,278	80.4%
Friends or wider family circle	283	1,184	19.2%	1,814,934	7,559,853	19.4%
Acquaintances	97	1,380	6.6%	372,523	9,002,264	4.0%
Strangers	90	1,387	6.1%	247,767	9,127,020	2.6%
Colleagues	59	1,418	4.0%	464,816	8,909,971	5.0%
Teacher/student	17	1,460	1.2%	153,811	9,220,976	1.6%
Unknown	12	1,465	0.8%	70,506	9,304,281	0.8%

**Note:* Because participants could self-report more than one relationship among interlocutors in the text, each row should be interpreted independently of one another.

Table 6. Communicative purpose(s) of LANA-CASE texts, adapted from Egbert et al., 2021 and Biber et al., 2021

Purpose	Texts tagged for this purpose	Texts <i>not</i> tagged for this purpose	% of texts tagged for this purpose	Words in texts tagged for this purpose	Words in texts <i>not</i> tagged for this purpose	% of words in texts tagged for this purpose
Sharing feelings and opinions	1,086	391	73.5%	7,737,746	1,637,041	82.5%

Joking around	954	823	64.6%	6,840,711	2,534,076	73.0%
Describing or explaining the past	875	602	59.2%	6,092,688	3,282,099	65.0%
Situation-dependent commentary	757	720	51.3%	4,741,954	4,632,833	50.6%
Figuring things out	683	794	46.2%	4,676,554	4,698,233	49.9%
Describing or explaining the future	634	843	42.9%	4,140,593	5,234,194	44.2%
Describing or explaining (time-neutral)	548	929	37.1%	3,898,923	5,475,864	41.6%
Giving advice and instructions	442	1,035	29.9%	3,312,267	6,062,520	35.3%
Engaging in conflict	119	1,358	8.1%	824,348	8,550,439	0.8%

**Note:* Because participants could self-report more than one communicative purpose of the text, each row should be interpreted independently of one another.

2.2. Demographic characteristics of LANA-CASE speakers

While Section 2.1 provides an overview of details related to LANA-CASE texts, the present section summarizes the demographic characteristics of the speakers who produced each text.

To be eligible to submit a recording to LANA-CASE, participants had to:

- speak English as a primary language;
- have lived in the U.S. since before elementary school; and
- be at least 18 years old (see Hanks et al., 2024, pp. 26-28).

In order to collect a corpus that was as balanced as possible across demographic groups, we recruited eligible participants according to pre-determined selection criteria, following the sampling methods of the demographically sampled portion of the British National Corpus (Leech, 1993). The selection criteria reflect the specific demographic groups that were targeted through recruitment (namely, with respect to their age, gender, geographic region, geographic setting, and race/ethnicity), from which we aimed to collect balanced data that are proportional

to the overall U.S. population. Recruitment methods were adapted iteratively throughout the compilation process to address limitations in the sample as much as possible (see Hanks et al., 2024, p. 28-36). While additional demographic information was collected from each speaker, such as education level and occupation (see Section 3.2 for an overview of LANA-CASE metadata), we did not attempt to balance corpus composition according to these descriptive criteria.

A summary of the composition of the corpus according to the selection criteria is found in the following tables.⁵ See Table 7 for a summary of the ages of LANA-CASE speakers, Table 8 for gender, Table 9 for geographic region, Table 10 for geographic setting, and Table 11 for race/ethnicity. In each table, the “Speakers” column provides the raw number of speakers in LANA-CASE who fall into each demographic category, and the “% Speakers” column provides the percentage of speakers in LANA-CASE who fall into each demographic category. The “Words” column provides the raw number of words that speakers from that demographic group contribute to the corpus,⁶ and the “% Words” column provides the percentage of words in LANA-CASE that speakers from that demographic group contribute to the corpus. Values were computed using a custom Python script. Because participants could self-report more than one race/ethnicity, the percentages in each row of Table 11 should not be summed. Rather, they should be interpreted independently of one another, where the numerator is the number of speakers (or words from these speakers) who identified as a race/ethnicity and the denominator is the total number of speakers (or words) in the corpus.

Table 7. Ages of LANA-CASE speakers

Age range	Speakers	% Speakers	Words	% Words
18-29	664	59.5%	5,992,549	64.0%

⁵ A summary of corpus composition according to descriptive criteria is available in the corpus manual (Hanks et al., under contract).

⁶ The total word count attributed to individual speakers in the corpus is slightly lower than the word count for the full corpus, since some words were spoken by multiple speakers and/or could not be attributed to a specific speaker with certainty.

30-39	132	11.8%	1,332,562	14.2%
40-49	93	8.3%	748,696	8.0%
50-59	89	7.9%	506,818	5.4%
60-69	51	4.6%	365,179	3.9%
70-79	33	3.0%	178,924	1.9%
80-102	51	4.6%	211,627	2.3%
Unknown	3	0.3%	28,732	0.3%
Total	1,116	100%	9,365,087	100%

Table 8. Genders of LANA-CASE speakers

Gender	Speakers	% Speakers	Words	% Words
Women	673	60.3%	5,668,040	60.5%
Men	363	32.5%	2,728,678	29.1%
Other	73	6.6%	930,616	10.0%
Unknown	7	0.6%	37,753	0.4%
Total	1,116	100%	9,365,087	100%

Table 9. Geographic regions of LANA-CASE speakers

Region	Speakers	% Speakers	Words	% Words
West	417	37.4%	3,080,669	32.8%
South	336	30.1%	2,930,809	31.3%
Midwest	189	16.9%	1,878,767	20.1%
Northeast	158	14.2%	1,411,022	15.1%
Not currently living in the U.S.	3	0.3%	6,393	0.1%
Unknown	13	1.1%	57,427	0.6%
Total	1,116	100%	9,365,087	100%

Note: Refer to Table 12 for a list of U.S. states and districts that map onto each geographic region.

Table 10. Geographic settings of LANA-CASE speakers

Setting	Speakers	% Speakers	Words	% Words
(Sub)urban	962	86.2%	8,142,210	86.9%
Rural	128	11.5%	1,138,540	12.2%
Unknown	26	2.3%	84,337	0.9%
Total	1,116	100%	9,365,087	100%

Table 11. Race/ethnicities of LANA-CASE speakers

Race/ethnicity	Speakers who identify as this race/ethnicity	Speakers who <i>do not</i> identify as this race/ethnicity	% of speakers who identify as this race/ethnicity	Words uttered by speakers who identify as this race/ethnicity	Words from speakers who <i>do not</i> identify as this race/ethnicity	% of words from speakers who identify as this race/ethnicity
White	860	256	77.1%	7,430,468	1,934,619	79.3%
Hispanic or Latino	142	974	12.7%	958,704	8,406,383	10.2%
Asian	96	1,020	8.6%	997,487	8,367,600	10.7%
Black or African American	86	1,030	7.7%	643,604	8,721,483	6.9%
American Indian or Alaska Native	41	1,075	1.1%	250,368	9,114,719	2.7%
Jewish	9	1,107	0.8%	163,259	9,201,828	1.7%
Middle Eastern	5	1,111	0.4%	79,844	9,285,243	0.9%
Native Hawaiian or Pacific Islander	8	1,108	0.7%	31,263	9,333,824	0.3%
Unknown	28	1,088	2.5%	97,354	9,267,733	1.0%

*Note: Because participants could self-report more than one race/ethnicity, each row should be interpreted independently of one another.

2.3. Evaluating the composition of LANA-CASE

Given the composition of LANA-CASE presented in the previous sub-sections, we encourage corpus users to draw on the situational characteristics of texts and demographic characteristics of speakers when evaluating LANA-CASE for whether and how to use it in their own research. Findings from a corpus sample can only be generalized to a target language domain to the extent that the language in the corpus represents that domain (Egbert, Biber & Gray, 2022). Researchers should evaluate LANA-CASE with respect to their target language domain and their specific research questions to determine whether, and to what extent, this corpus can be useful and appropriately generalizable. In line with this, we invite users to sub-sample from the corpus as necessary in order to address their particular research questions and/or improve comparability with other corpora (e.g., comparability with the Spoken BNC2014; Love et al., 2017).

Hanks, Egbert, Biber, Larsson, and McEnery (under review) provide a general domain evaluation of the corpus for the degree to which it is representative of the domain of everyday, private conversation in American English. This evaluation follows the framework for corpus design presented in Egbert et al. (2022) and discusses implications of corpus composition for the research process. This evaluation highlights both strengths and limitations of the representation of LANA-CASE. Overall, Hanks et al. (under review) reveal that many of the situational characteristics of corpus texts align closely with the target domain (i.e., corpus texts take place in a shared/synchronous context, and many corpus texts are unplanned and unedited, are spoken face-to-face, have a largely phatic goal, focus on *sharing feelings and opinions*, and do not focus on *giving advice and instructions*). In addition, the representation of geographic regions, geographic settings, and race/ethnicities of LANA-CASE speakers aligns relatively closely with the target domain of proportional representation of diverse speakers across the U.S. On the other hand, some constraints of corpus composition include (a) many texts that incorporate *figuring things out* and *joking around* as only minor (rather than major) purposes, and (b) some over- and under-sampling of speakers from different age ranges and gender identities.

We hope that the corpus description in this paper, combined with the general domain evaluation provided in Hanks et al. (under review), will provide researchers with a useful starting point for their own evaluations of the corpus with respect to their particular target language domains and research questions, allowing them to determine how far generalizations from the results may be drawn.

While this section has focused on the composition of LANA-CASE in terms of its texts and speakers, the following section provides logistical guidance for accessing and using the corpus.

3. Accessing and using the corpus

LANA-CASE is available open access on #LancsBox X (<https://lancsbox.lancs.ac.uk/>; Brezina et al., 2015). In addition, the corpus files are available to download in different formats: i) plain text and ii) XML. Plain text files can be downloaded with transcription tags (see Figures 1 and 2) and without transcription tags (see Figure 3) from <https://cass.lancs.ac.uk/lana-case/>. All plain text files are accompanied by two metadata CSV files, one which provides information about speakers in the corpus, and another which provides information about the texts in the corpus. Plain text versions of the corpus may be most suitable for researchers interested in analyzing corpus texts using their own custom-made scripts or qualitatively analyzing longer stretches of conversation by reading corpus texts. XML files (with integrated metadata) can also be downloaded (see Figure 4). These files are tagged using the C5 tagset from CLAWS (Garside & Smith, 1997), the same tagger used in the Spoken British National Corpus 2014 (Love et al., 2017).

Figure 1. Example of a plain text metadata header

```

<file_name: 0042_LANACASE_f2f_2S_20min>
<submission_date: oct 8th 2024>
<submission_state: rhode island>
<duration_minutes: 20>
<duration_hours: 0.33>
<number_of_participants: 2>
<speaker_ids: NP 1214? MANY 1215 1214>
<mode: f2f>
<setting_standardized: home, assisted living center>
<setting_self-reported: assisted living center>
<relationship_standardized: friends or wider family circle>
<relationship_self-reported: friends or wider family circle ; 1214 and 1215 are
female residents at the same assisted living center>
<communicative_functions: unknown>
<references_to_recording: none recorded>
<recording_device: odyssey digital recorder>
<transcriber: 04>
<moderation: yes>
<participant_comments: none>
<team_comments: friends talking about friends, family, their lives, and dogs>

```

Figure 2. Example of a plain text file with transcription tags

```

<1215>: {whisper=it was freezing} the deck is cold
<1214>: [00:01:10] {name} mentioned that
<1215>: [00:01:12] really cold
<1214>: [00:01:14] [OL] yeah
<1215>: [00:01:14] then I went
<1214>: [OL] {u=?}
<1215>: downstairs and sat outside and it was fine I had this on for a minute then
I took it off but it was still it was chilly the wind was blowing but it was
[e=mi|sc]
<1214>: yeah
<1215>: I thought about having a blanket and curling up on the grass [.] and then
I thought about all the bugs coming to bite me
<1214>: [00:01:35] it'd be just your luck
<1215>: (laugh) {u=and I}
<1214>: [OL] and sit it on an anthill
<1215>: [00:01:39] I know
<1214>: (laugh)
<1215>: I know I know then I thought no it sounds lovely but it's really not a
good idea to do that {u=so}

```

Figure 3. Example of a plain text file without transcription tags

```

it was freezing the deck is cold
mentioned that
really cold
yeah
then I went
downstairs and sat outside and it was fine I had this on for a minute then I took
it off but it was still it was chilly the wind was blowing but it was
yeah
I thought about having a blanket and curling up on the grass and then I thought
about all the bugs coming to bite me
it'd be just your luck
and I
and sit it on an anthill
I know
I know I know then I thought no it sounds lovely but it's really not a good idea
to do that so

```

Figure 4. Example of an XML file

```

<u age="83" ageEstimated="False" ageRange="80-89" ageRange1994="60+"
birthYear="1942" bornUSA="True" confident="True" education="Graduate degree"
emigUSAPreElem="None" gender="Female" genderSelfRep="Female (queer)"
geogSettCmnts="None" geogSettStand="urban or suburban" idx="19"
langSpknAtHome="English" langSpknOutsdHome="English" multiSpk="False"
occupSelfRep="retired social worker professor, psychotherapist (long time),
ACTIVIST" occupation="retired,teaching,research,medicine" overlap="False"
raceEthnSelfRep="White,Jewish" raceEthnStand="Jewish" redacted="False"
region="Northeast" spkID="1215" stateCurrent="Rhode Island"
stateMajority="Connecticut" stateStand="Connecticut">
  <s>
    <shift position="start" type="whisper"/>
    <w pos="PPH1" sem="Z8">
      it
    </w>
    <w pos="VBDZ" sem="Z5">
      was
    </w>
    <w pos="VVG%" sem="O4.6">
      freezing
    </w>
    <shift position="end" type="whisper"/>
    <w pos="AT" sem="Z5">
      the
    </w>
    <w pos="NN1" sem="M4">
      deck
    </w>
    <w pos="VBZ" sem="A3">
      is
    </w>
    <w pos="JJ" sem="S1.2.1">
      cold
    </w>
  </s>
</u>

```

The following sub-sections provide an overview of the transcription conventions (Section 3.1), the metadata file for texts (Section 3.2), and the metadata file for speakers (Section 3.3) (see the corpus manual for more detailed information; Hanks et al., in preparation).

3.1. Transcription conventions

LANA-CASE transcription conventions were adapted from those used in the Spoken British National Corpus 2014 (Love et al., 2017). Utterances were transcribed precisely as they were heard, including any potential grammatical mistakes, false starts, overlapping speech, laughter, and (non)linguistic vocalizations. Beyond this, transcription tags were incorporated to record metadata and speaker IDs, nonlinguistic vocalizations, stylized utterances, and contextual

information. Texts can be downloaded with or without these additional transcription tags (see Figures 2 and 3, respectively). Details about the transcription tags are provided below.

Metadata and speaker IDs

Metadata and speaker IDs are annotated with <angle brackets>. The header includes metadata about the text (see Figure 1; this metadata can also be found in the metadata file for texts; see Section 3.2). Throughout the transcript, speaker IDs are recorded in angle brackets at the beginning of each line (see Figure 2).

Nonlinguistic vocalizations

Nonlinguistic vocalizations such as laughter are transcribed within (parentheses). They are transcribed according to a closed list of items, which includes the following:

- (laugh)
- (gasp)
- (sigh)
- (whistle)
- (hum)
- (yell), for wordless yells or screams
- (sob), for wordless crying
- (mimic=OBJECT OF MIMIC), for imitating sounds; replace OBJECT OF MIMIC with the object being mimicked such as an animal or machinery

Stylized utterances

Stylized utterances such as singing are transcribed within {curvy brackets}. They are transcribed according to a closed list of items, which includes the following, in which WORDS are replaced with the words that have been uttered (or replaced with “?” if they are unknown):

- {u=WORDS}, for words that are unclear but have been guessed by the transcriber
- {sing=WORDS}
- {whisper=WORDS}
- {f=LANGUAGE=WORDS}, where LANGUAGE is replaced with the foreign language that is spoken; ? is used when the transcriber is unsure

The anonymized names of people and places uttered in the conversation are enclosed within curly brackets, such as in {name} and {place}.

Contextual information

Contextual information—including pauses, overlapping speech, timestamps, and sounds in the environment—are annotated with [square brackets]. Short pauses between 1 and 5 seconds are presented as [.] , while pauses that are five seconds long or longer are presented as [...]. When the utterance of one speaker begins before the previous speaker’s utterance has come to a close, this overlap is indicated by the [OL] tag at the start of the new speaker’s turn. Timestamps are annotated as [hour:minute:second] (e.g., [00:20:52]) at regular intervals, although not necessarily at the start of each turn. Utterances spoken by ineligible speakers are not transcribed; they are instead replaced with the tag [NP] (for “non-participant”).

Sounds in the environment are transcribed only when the sound is relevant to the conversation, affecting in some way the course of the conversation, such as when a speaker utters “it’s my sister” in response to the sound of a phone ringing. These sound events are annotated as [e=SOMETHING], where SOMETHING is replaced with a description of the sound. Some examples include the following:

[e=music]

[e=electronic media], such as sound from a TV

[e=background talk]

[e=Siri responds]

3.2. Metadata for texts

The metadata file for texts contains information provided by participants about each text in the corpus. This metadata is also available at the top of each corpus text downloaded with transcription tags (see Figure 2). A brief overview of each column in the spreadsheet is provided below.

- A. File name: This consists of a four-digit text ID, followed by “LANACASE”, the mode (“f2f” or “remote”), number of speakers (e.g., “2S,” “3S,” “4S”), and the audio duration in minutes (e.g., “14min,” “135min”). For example, the first text in the corpus is named 0001_LANACASE_f2f_2S_38min.
- B. XML ID: This provides information about the text’s corresponding XML ID.
- C. XML staticID: This provides information about the text’s corresponding XML staticID.
- D. Submission date: The date when the text was submitted. This date does not necessarily align with the date of recording, as the submission guidelines did not require that texts be submitted immediately after recording.
- E. Submission state: The state where the text was submitted. This state does not necessarily align with any of the speakers’ standardized states (see Section 3.3) nor the state of recording, as the submission guidelines did not require that texts be submitted immediately after recording.
- F. Duration_minutes: Total length of the recording in minutes.
- G. Duration_hours: Total length of the recording in hours.
- H. Number of participants: Total number of registered participants in the text.
- I. Speakers included: The speaker ID of every registered speaker in the text is listed here.
- J. Mode: Responses include “f2f” (face-to-face) or “remote” (e.g., via Zoom, online audio/video chat through Discord, or phone).
- K. Setting_standardized: The setting of the text has been categorized as one of the following options: “Private,” “Semi-private,” and “Public.” For example, a self-reported setting of ‘at home’ was categorized as “Private”, ‘in a classroom’ was categorized as “Semi-private”, and ‘on a bus’ was categorized as “Public”. More details about the standardization procedure are available in the corpus manual (Hanks et al., in preparation).
- L. Setting_self-reported: The setting of the text is retained exactly as reported by participants.
- M. Relationship_standardized: Responses include one or more of the following: “Close family, partners, or very close friends,” “Friends or wider family circle,” “Teacher/student,” “Colleagues,” “Acquaintances,” “Strangers,” or “Other.”

- N. Relationship_self-reported: More precise relationship distinctions that were recorded by participants are listed here (e.g., “good friends since 6th grade,” “roommates,” “161 is an alum from 155's program and has been teaching in Japan for the past six or seven years. He was 155's student”).
- O. Communicative function(s): Responses include one or more of the following (adapted from a list of communicative purposes of conversation developed by Egbert et al., 2021 and Biber et al., 2021): “Describing or explaining the past,” “Describing or explaining the future,” “Describing or explaining (time neutral; e.g., explaining the difference between two products),” “Giving advice and instructions,” “Sharing feelings and evaluations/opinions,” “Figuring things out (e.g., making plans or solving problems),” “Engaging in conflict,” “Joking around,” and “Commenting about your immediate surroundings (e.g., discussing a board game you're playing).”
- P. References to recording: Timestamps marking locations in the text where speakers referenced the fact that they are recording (if any) are listed here. These timestamps may be used as one metric for determining the extent to which participants were aware of the fact that their conversation was being recorded, which may have an impact on the authenticity of the language exchanged.
- Q. Recording device: The description of the audio device used to record the audio file is retained exactly as recorded by participants.
- R. Transcriber: This column provides the transcriber ID(s).
- S. Moderation: Information about whether the transcription was moderated by another transcriber or volunteer is provided as “yes” or “no”.
- T. Participant comments: Additional comments have been anonymized and retained exactly as recorded by participants. These comments often focus on participants in the recording (e.g., “Baby in recording is 9 months old. She is not speaking yet.”), the relationship of participants (e.g., “I don't see my grandma often”), potential issues with the audio quality (e.g., “This recording was a pretty candid interaction between me and my ex partner after a recent breakup. Apologies for the crying, I hope the audio is mostly clear.”), additional context (e.g., “By the end we were considerably drunk”), or reflections about the process (e.g., “It was incredibly enjoyable to relax for an hour and have a productive conversation. The time seemingly flew by!”).

- U. Team comments: Additional comments provided by the team and/or transcribers are listed here. These comments often focus on topics in the recording (e.g., “waiting in truck for rain to stop so can go inside, attempt to exercise in truck”), the quality of the recording (e.g., “a lot of background noise”), or extenuating circumstances about participants (e.g., “1070 has a strong Spanish accent and is often unclear”).
- V. Word count_inclusive: This is the inclusive word count for the text (which includes unclear and anonymized speech and considers merged forms such as “can’t” and “gonna” as multiple words).
- W. Word count_conservative: This is the conservative word count for the text (which does not include unclear or anonymized speech and considers a word to be any string of characters separated by white space).

3.3. Metadata for speakers

The metadata file for speakers contains information about each speaker in the corpus, including details about selection criteria (i.e., age, gender, geographic region, geographic setting, race/ethnicity; discussed in Section 2.2) and the other metadata collected (e.g., education level, occupation). A brief overview of each column in the metadata spreadsheet is provided below.

- A. Speaker ID: These are four-digit numbers that are unique to each participant, sorted in ascending order.
- B. Were you born in the United States?: Responses include “yes” or “no.”
- C. Did you move to the United States before you started elementary school?: Only recordings in which all speakers responded “yes” or “N/A” to this question were included in the corpus.
- D. Birth year: Participants’ self-reported birth year has been standardized as four-digit birth years (e.g., “1972”).
- E. Age: Participants’ age has been calculated by subtracting their self-reported birth year from the date of submission.
- F. Gender_standardized: Participants’ gender has been categorized as one of three options: “Male,” “Female,” or “Other.”

- G. Gender_self-reported: Participants' self-reported gender is retained exactly as recorded by participants. Responses include "Male," "Female," and a wide range of open-ended responses for other genders.
- H. Race/ethnicity_standardized: Participants' race/ethnicity has been categorized as one or more of the following: "American Indian or Alaska Native," "Asian," "Black or African American," "Hispanic or Latino," "Jewish," "Middle Eastern," "Native Hawaiian or Pacific Islander," "White", or "Other."
- I. Race/ethnicity_self-reported: Participants' self-reported race/ethnicity is retained exactly as recorded by participants. Responses include one or more of the following: "American Indian or Alaska Native," "Asian," "Black or African American," "Hispanic or Latino," "Native Hawaiian or Pacific Islander," "White", and a wide range of open-ended responses for other race/ethnicities.
- J. What language(s) do you speak at home?: Participants' self-reported language(s) is/are retained exactly as recorded by participants.
- K. What language(s) do you speak outside of the home?: Participants' self-reported language(s) is/are retained exactly as recorded by participants.
- L. What is the highest level of education you have completed?: Responses include "Less than high school," "High school graduate," "Trade school certificate (e.g., electrician, commercial driver, cosmetology, etc.)," "Undergraduate degree," and "Graduate degree."
- M. Occupation_ standardized: Participants' occupation has been categorized into groups (see Hanks et al., in preparation for more details).
- N. Occupation_self-reported: Participants' self-reported occupation is retained exactly as recorded by participants.
- O. Geographic setting_standardized: Responses include "urban or suburban" or "rural."
- P. Geographic setting_comments: Additional comments about participants' setting, if any, are retained exactly as recorded by participants, except for cases in which the participant recorded the precise name of the city or town where they live.
- Q. State_standardized: This is the state where participants have lived for more than half their lives ("State_majority" in Column S), or, if participants have not lived in one state for more than half their lives, this is their current state of residence ("State_current" in Column R).

- R. **State_current**: This is the state in which participants currently live. Responses include all 50 states, Washington D.C., and “not currently living in the U.S.” (an option added to accommodate students on study abroad and similar situations).
- S. **State_majority**: If participants have lived in one state for more than half of their lives, this state is listed here. Responses include all 50 states, Washington D.C., and “N/A.”
- T. **Region**: Region was recorded based on participants’ standardized state (Column Q), according to categories used in the U.S. Census, shown in Table 12.
- U. **Number of files**: This is the number of texts in which the speaker is involved.
- V. **File names**: These are the names of the texts in which the speaker is involved.
- W. **Number of words_inclusive**: This is the inclusive word count attributed to this speaker (which includes unclear and anonymized speech and considers merged forms such as “can’t” and “gonna” as multiple words).
- X. **Number of words_conservative**: This is the conservative word count attributed to this speaker (which does not include unclear or anonymized speech and considers a word to be any string of characters separated by white space).

Table 12. U.S. states included in each region

West	Midwest	Northeast	South
Alaska	Illinois	Connecticut	Alabama
Arizona	Indiana	Maine	Arkansas
California	Iowa	Massachusetts	Delaware
Colorado	Kansas	New Hampshire	Florida
Hawaii	Michigan	New Jersey	Georgia
Idaho	Minnesota	New York	Kentucky
Montana	Missouri	Pennsylvania	Louisiana
Nevada	Nebraska	Rhode Island	Maryland
New Mexico	North Dakota	Vermont	Mississippi
Oregon	Ohio		North Carolina
Utah	South Dakota		Oklahoma
Washington	Wisconsin		South Carolina
Wyoming			Tennessee

			Texas
			Virginia
			Washington D.C.
			West Virginia

4. Conclusion

LANA-CASE (a) represents everyday, private conversation; (b) represents diverse speakers of American English across the U.S.; (c) is large in size (i.e., contains several millions of words); (d) represents conversational language from 2022-2024; and (e) is freely available. This is the first corpus which exhibits all five of these characteristics, which we hope will enable a wide range of investigations into this particular register and variety of English, addressing research questions that have not previously been accessible to us as a field. LANA-CASE consists of 9,374,787 words (1,330 hours) of conversation from over one thousand speakers across the U.S. It exhibits a wide range of situational characteristics in its texts as well as demographic characteristics in its speakers. In this paper, we have presented key information that users will need to understand the corpus composition as well as to access and use the corpus and metadata files. We hope that this corpus, the culmination of our efforts over the past three years, will prove to be a useful resource for research on various aspects of American English conversation as well as a pedagogical tool for classroom activities and assignments.

Acknowledgements

We are grateful to *Lancaster University's Global Advancement Fund, UK Research and Innovation (MR/V022954/1), Northern Arizona University's Faculty Course-based Undergraduate Research Experience Development Grant, the Northern Arizona University Corpus Lab, and Northern Arizona University's SGS Award*, whose support has made this project possible. We also gratefully acknowledge the participants who contributed recordings and our collaborators who helped with recruitment, transcription, and data management: Eric Audrey Christensen, Cansu Avci, Ruth Avon, Helen Baker, Theodora Bofman, Alexander Brogan, Earl Brown, Landon Bryant, Amy Buck, Marcella Caprario, Lydia Cooper, Nur Yağmur Demir, Deeya Dhanak, Dilara

Dikilitas, Daniel Dixon, Tülay Dixon, Gözde Durgut, Idée Edalatishams, Michael Edens, Faith Ennis, Louise Eriksson, Ella Fearn, Emilia Fletcher, Macy Elizabeth Floyd, Rhona Gardner, Kevin Gerigk, Larissa Goulart, Marianna Gracheva, Heidi Guenther, Tilly Hallett, Michael Hanks, Brett Hashimoto, Daniel W. Hieber, Kevin Hirschi, Samuel Hollands, Yuhang Hu, Dan Isbell, Yejin Jung, Daniel Keller, Kelly Kendro, Taehyeong Kim, Elizaveta Kuznetsova, Kris Kyle, Febriana Lestari, Zach Marlin, Lia Martin, Kim Thien Nguyen, Sadie Ossanlow, Laylah Richards, Michelle Richter, Ellie Robins, Seren Rutter, Juliana Ser, Emily Sharp, Jon Smart, Jordan Smith, Rebecca Spencer, Anne Stoughton, Hanady Suleiman, Iia Vlasova, Elena Walker, Emma Winn, Shelley Yang, Katherine Yaw, Shujing Zhao, HuHot Mongolian Grill Flagstaff, Program in Intensive English at Northern Arizona University, and Assisted Living Centers across the US.

References

- Baker, P. (2017). *American and British English: Divided by a common language?*. Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. Longman.
- Biber, D., Egbert, J., Keller, D., & Wizner, S. (2021). Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics*, 171, 20–35.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocational networks. *International Journal of Corpus Linguistics*, 20(2). 139–173.
- Burnard, L. (2002). Where did we go wrong? A retrospective look at the British National Corpus. In B. Kettemann & G. Markus (Eds.), *Teaching and learning by doing corpus analysis* (pp. 51–70). Brill.
- Coats, S. (2023). Dialect corpora from YouTube. In B. Busse, N. Dumrukic, & I. Kleiber (Eds.), *Language and linguistics in a complex world*, 79–102. De Gruyter.
<https://doi.org/10.1515/9783111017433-005>
- Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159–190. <https://doi.org/10.1075/ijcl.14.2.02dav>

- Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000-2005). Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.
- Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press.
- Egbert, J., Biber, D., Gray, B., & Larsson, T. (2025). Achieving stability in corpus-based analysis of word types. *International Journal of Corpus Linguistics*.
- Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41(5-6), 715-737.
- Garside, R. & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102-121). Longman.
- Hanks, E., McEnery, T., Egbert, J., Larsson, T., Biber, D., Reppen, R., Baker, P., Brezina, V., Brookes, G., Clarke, I., & Bottini, R. (2024). Building a spoken corpus of conversational American English: Challenges and innovations in corpus design. *Research in Corpus Linguistics*, 12(2), 24-44. <https://doi.org/10.32714/ricl.12.02.03>
- Hanks, E., Egbert, J., Biber, D., Larsson, T., & McEnery, T. (under review). Designing and evaluating the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE).
- Hanks, E., McEnery, T., Egbert, J., & Larsson, T. (in preparation). *Corpus manual for the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE)*.
- Kendall, T. & Farrington, C. (2023). *The Corpus of Regional African American Language*. Version 2023.06. Eugene, OR: The Online Resources for African American Language Project. <https://doi.org/10.7264/1ad5-6t35>.
- Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2021). Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh. *Language Resources and Evaluation*, 55. 789-816.
- Larsson, T., Biber, D., Kim, T., & Holmberg, A. J. (2026). Is casual conversation linguistically similar to other spoken registers?
- Leech, G. (1993). 100 million words of English. *English Today* 9(1), 9-15.

- Love, R. (2020). *Overcoming challenges in corpus construction: The Spoken British National Corpus 2014*. Routledge.
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344. <https://doi.org/10.1075/ijcl.22.3.02lov>
- Nofsinger, R. E. (1991). *Everyday conversation*. Sage Publications.
- Pöldvere, N., Johansson, V., & Paradis, C. (2021). On the London–Lund Corpus 2: Design, challenges and innovations. *English Language & Linguistics*, 25(3), 459-483.
- Shirk, J., Ballard, H., Wilderman, C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B., Krasny, M., & Bonney, R. (2012). Public participation in scientific research: A framework for deliberate design. *Ecology and Society*, 17(2), 1–20.
- Stern, K. (2005). *The Longman Spoken American Corpus: Providing an in-depth analysis of everyday English*. Pearson Education.
- Svartvik, J. L. & Quirk, R. (1980). *A corpus of English conversation*. C.W.K. Gleerup.
- U.S. Census Bureau. n.d. *Explore census data*. <https://data.census.gov/> (June 2022).