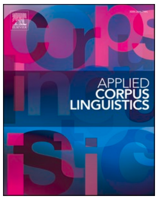
Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp



The DAIS-C: A small, specialised, spoken, schizophrenia corpus

Oliver Delgaram-Nejad [a,*], Dawn Archer [b], Gerasimos Chatzidamianos [c], Louise Robinson [d], Alex Bartha [e]

[a] *Department of Linguistics, Department of Psychology, Manchester Metropolitan University, Manchester M15 6LL, UK*
[b] *Manchester Metropolitan University, UK*
[c] *National Institute for Deaf People, Manchester Metropolitan University, Greece*
[d] *Lancashire and South Cumbria NHS Foundation Trust, University of Manchester, UK*
[e] *East London NHS Foundation Trust, UK*

ARTICLE INFO

Keywords:
Corpus
Schizophrenia
Formal thought disorder
Creativity
Resource

ABSTRACT

This paper describes the design and development of the DAIS-C (Discussing Abstract Ideas in Schizophrenia Corpus), a small, specialised corpus of spoken language in which speakers with a diagnosis of schizophrenia and those with no self-reported psychiatric or neuroleptic history were interviewed on the same topics. The corpus was constructed to allow for comparative analyses of speech behaviour in relation to linguistic creativity and formal thought disorder (FTD), but additional steps were taken to ensure that the corpus could be of use to other researchers and research questions. The present paper covers design decisions relevant to the construction of clinical corpora alongside information about the corpus of potential use to researchers interested in its use.

The present paper describes the design and characteristics of a corpus built to support an investigation of linguistic creativity and formal thought disorder (FTD) in schizophrenia. A key feature of its design was its capacity for reuse by other researchers, which is the focus of this paper. It is intended to act both as a description of the corpus' content and as a reference aid for potential users.

Although corpus linguists make up the primary audience of this paper, introductory sections on building spoken and specialised corpora are provided for interdisciplinary readers, unfamiliar with corpus linguistics, who may also be interested in using the corpus and/or building corpora of their own. These sections are also included given their relevance to the design criteria discussed later.

This paper also functions as a call for more linguistic work within the medical humanities to consider the theoretical aspects of corpus design. These guidelines are practically useful and also empirically important, and yet many text analytic studies of clinical populations that employ corpus methods tend not to incorporate corpus building practices into data collection. More awareness of corpus design will lead to improved data quality and greater confidence in the findings reported by these studies.

The paper begins by outlining reasons for corpus construction and related guidance. It then discusses decisions specific to the development of a corpus interested in clinical questions, particularly the symptomatology of schizophrenia and how these considerations were integrated into extant guidance and theory on corpus design. The final part presents a brief review of the resulting corpus' characteristics, which were affected by design issues secondary to recruitment challenges/the COVID pandemic. These are explored ahead of a discussion of potential applications of the corpus.

Background

Why build a corpus

Corpus linguistics is gaining popularity as a research method, in and outside of linguistics itself (see Mouritsen 2019 for practical applications in law; Mitkov 2022). This has led to an increase in the production of reference and specialised corpora. Reference corpora tend to be larger and aim to represent language varieties as a whole, such as the British National Corpus (BNC, 2007) or Corpus of Contemporary American English (COCA; Davies, 2015), whereas specialised corpora focus on specific linguistic contexts and communities.

Researchers wanting to use corpora to answer a research question will need to check for existing corpora or build one if nothing suitable exists. Corpus design stems from the original research question. A corpus is not just a text database. It is a body of linguistic examples curated to

* Corresponding author.
E-mail address: olinejad@protonmail.com (O. Delgaram-Nejad).

https://doi.org/10.1016/j.acorp.2023.100069
Received 16 April 2023; Received in revised form 11 August 2023; Accepted 19 August 2023
Available online 23 August 2023

answer a set question (Randi, 2010).

There is no quantitative answer to the question of how big a corpus should be. Representativeness describes the extent to which a corpus accounts for the language variety it samples. A corpus of an author's collected works, for example, would be completely representative. This is impractical in most cases, and enough data for an accurate representation usually suffices (Adolphs and Knight, 2010).

Representativeness is more challenging with respect to schizophrenia. One reason is symptom heterogeneity (Oomen et al., 2022). Schizophrenia symptoms range in nature and degree and affect linguistic production and comprehension (McKenna and Oh, 2005). It is arguable that we cannot currently assess representativeness in this population because the true extent of linguistic variation is not yet understood (McKenna and Oh, 2005; Mikesell and Bromley, 2016).

This paper discusses the design and characteristics of the DAIS-C (Discussing Abstract Ideas in Schizophrenia Corpus), which was built to answer the following question*: is there a relationship between linguistic creativity and formal thought disorder in schizophrenia*?

Despite an increase in corpus linguistic applications, no British English corpus of speech in schizophrenia prior to the DAIS-C existed. A reference schizophrenia corpus would be a phenomenal undertaking, one exceeding the scope of a thesis. A specialised corpus built to explore linguistic creativity in this population offers a useful first step, nonetheless. The next sections review best practices in the design of small, specialised, and spoken corpora. These are synthesised to form a set of requirements for the DAIS-C. A description of how I approached these requirements follows before a summary of the DAIS-C's main characteristics.

Building spoken language corpora

Spoken language corpora fall under the class of special corpora, meaning that they do not necessarily seek to represent the full extent of a language variety but rather a special case of language use. Prominent examples like the BNC spoken represent speech orthographically and sample from a range of spoken contexts such as lectures, speeches, and conversations. More recent work has taken interest in informal conversation (CANCODE; McCarthy 1998).

FTD manifests in speech and writing, but more work has sampled FTD in spoken contexts. Historic work has noted that FTD in speech is more readily elicited in the context of proverb interpretation tasks and comprehension subsets of standardised intelligence tests (Marengo et al., 1986). These approaches have been replicated substantially. Indeed, much of this work was reviewed systematically in Delgaram-Nejad et al. (2020). It is important to point out that although interactional, these remain only semi-naturalistic events due to their location within the context of formal testing. Less work has examined FTD in the context of fluid, informal conversation (Mikesell and Bromley, 2016).

Several best practices exist for the construction of spoken corpora. There are both general construction guidelines and guidance specific to particular construction stages. Sinclair (2005) provides some of the most formal, comprehensive, and general (reproduced here from Adolphs and Knight 2010, p.39):

1 The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
2 Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.
3 Only those components of corpora which have been designed to be independently contrastive should be contrasted.
4 Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5 Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
6 Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.
7 The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
8 The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.
9 A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

It is recognised that complete adherence to Sinclair's guidelines is challenging in practice. There is good agreement, however, that they function well as guiding ideals (Adolphs and Knight, 2010). A corpus design that makes reasonable attempts to follow this advice as closely as possible stands a better chance of being reflective of the language variety under study and usable to the wider research community.

The importance of metadata is also stressed, and it is helpful to consider collecting editorial, analytic, descriptive, and administrative types of metadata (Burnard, 2005). 'Editorial' metadata provides information about how corpus components relate to original sources. 'Analytic' provides information about interpretation and analysis. 'Descriptive' provides classification data on internal and external properties. 'Administrative' provides information about the corpus itself, such as title, revisions, etc.

The ethics of spoken corpus construction require careful thought. Informed consent, despite being essential, should not only specify consent to record but also to distribute (Leech et al., 2014; Thompson, 2005). Anonymisation also requires care. Data that could potentially identify a participant must be located and obscured (Du Bois, 1992), and not all identifying features are immediately obvious. 'Raw' audio records may contain unique phonetic features that can potentially identify individuals (Adolphs and Knight, 2010). Anonymisation may also extend to sensitive topics (Wray et al., 1998).

Guidelines on audio recording in the construction of spoken corpora emphasise audio quality alongside an adequate account of the environmental features of a spoken interaction (Strassel and Cole, 2006). The transcription of spoken language is a complex task. Spoken language is fundamentally multimodal, with meaning constructed from textual, prosodic, gestural, and environmental elements (Adolphs and Knight, 2010). Representing this interplay in writing can be challenging, and investigators can quickly become consumed by attempts to capture the full richness of the data (Cook, 1990; McCarthy 1998; Carter, 2015; Halliday, 2004). Transcription ultimately boils down to theorising (Ochs, 1979; Edwards, 1993; Thompson 2005), and there is significant tension between validity and reading ease (Graddol et al., 1994).

There is a growing interest in and a need for spoken language corpora that deal with naturalistic interactions (Batinić et al., 2021). FTD can be elicited in informal spoken contexts, particularly when the discussion focuses on open, abstract topics. Best practices for the construction of spoken language corpora emphasise careful and systematic corpus construction, consideration of technical and environmental factors relevant to spoken discourse, collecting comprehensive metadata, practising ethical corpus construction, and transcribing on a robust theoretical basis. The next section discusses guidance on the development of small, specialised corpora.

Building small, specialised corpora

Specialised corpora are gaining popularity (Flowerdew, 2014) and represent a departure from the established trend of compiling sizable reference corpora. These smaller corpora focus on specific genres and registers.

Small corpora are unsuitable for some analyses because not all linguistic features manifest in small samples. Some lexicographical features are so rare that only a few examples appear in corpora composed of hundreds of millions of words. Grammatical patterns do however occur with enough regularity for reliable analysis within small corpora (Carter and Mncarthy, 1995). Smaller corpora also allow analysts to be more precise about the original contexts of use, because there tends to be less contextual variation (Flowerdew, 2004). Builders and analysts of small corpora are usually one and the same, and familiarity with the context allows analysts to supplement their quantitative observations with supportive qualitative analyses (Flowerdew, 2004; O'Keefe, 2007).

The present investigation is interested not only in schizophrenia and FTD but also the relationship between linguistic disturbances and their interactional contexts. Specialised corpora allow for a deeper examination of this context and the potential to build contextual variation into the design. A specialised corpus can be specialised in several ways: purpose of investigation (what), contextualisation (where, who, why), genre, type of text (conversation), subject matter, and variety of English (Flowerdew, 2004). They can also contain specialised sub-corpora, such as in the Hong Kong Corpus of Spoken English (Warren, 2004) that contains conversation, business, academic, and public sub-corpora. It is also recommended that builders of spoken, specialised corpora transcribe laughter and related features where the objective is to analyse interaction (Almut, 2010).

Even builders of specialised corpora are required to factor representativeness into their design. This has been defined as the extent to which the full range of variability is captured by the sample (Biber, 1993), with variability expressed as being either situational or linguistic (Biber, 1993; Almut, 2010). Situational variability refers to the spread of registers or genres in the population, whilst linguistic variability refers to the extent of linguistic variety in the population. It is argued that situational representativeness must be defined first to allow for the subsequent analysis of linguistic representativeness: the main thing is to ensure that samples are taken from a range of typical situations (Almut, 2010). Linguistic representativeness can be achieved with samples of 1000 words, and genres or registers can be well represented with samples as low as five in some cases, although ten is preferable according to Biber (1990).

Authors building specialised corpora for reuse by others can factor future use into their designs. Transcription conventions for specialised corpora tend toward 'one-offness' or the tendency for annotation to cater only to the needs of a given project (Almut, 2010). Planning for reuse by others can (and, where possible, arguably should) inform following design stages.

Larger corpora can also be used to support specialised corpus work, such as by checking whether high frequency words in the specialised corpus are more or less frequent in general usage (Almut, 2010). Specialised corpora therefore need not exist in a vacuum.

Interest in specialised corpora is increasing, especially among those interested in the role of context in interaction. Although not suitable for all analyses, they are well positioned for analysts interested in a close examination of features that appear reliably in small samples. They also suit analysts with a disposition toward mixed methods. Linguistic examinations of schizophrenia and FTD stand to benefit from a specialised corpus approach because general reference corpora do not adequately represent schizophrenia populations (Gabrić et al., 2021). Representativeness remains a consideration, but this can be partly addressed by sampling from a range of situational contexts. The next section brings together the guidance for building small, specialised corpora and spoken corpora as a set of operational requirements for the DAIS-C.

Requirements for the DAIS-C

The DAIS-C needed to permit an investigation of linguistic creativity, schizophrenia, and FTD. None of these concepts are particularly well defined *linguistically*. Requirements relevant to the research question are outlined below:

- Allows for linguistic creativity and FTD comparison.
- Compares schizophrenia and nonpsychiatric cohorts.
- Compares semi-naturalistic (experimental) and naturalistic (conversational) contexts.

Guidelines on the creation of spoken and specialised corpora were also important to the design. The spoken corpus requirements for the DAIS-C can be summarised as follows below:

- Close compliance with Sinclair's (2005) and Adolphs and Knight (2010, p.39) general guidance.
- Good audio quality and transcription of relevant environmental features.
- Comprehensive metadata covering editorial, analytic, descriptive, and administrative dimensions.
- Informed consent to record and also archive data *via* a repository for use by other researchers, comprehensive anonymisation, the avoidance of sensitive topics unless agreed by the participant, and, if applicable, the destruction of raw audio.
- Detailed and relevant transcription that captures key textual, prosodic, gestural, and environmental elements while preserving reading ease.

The specialised corpus requirements for the DAIS-C can be summarised thusly:

- Samples from a range of linguistic and situational contexts within the population.
- Gathers detailed contextual information.
- Aims for a minimum of 1000 words per speaker.
- Aims for a minimum of five samples per register or genre.
- Builds potential reuse by others into the planning and design.

A corpus that allows for an exploration of both the FTD framework (a theory-driven model of FTD that focuses on grammatical, word selection, thought completion, and discourse tracking features, under review) and linguistic creativity stands to benefit from a combination of best practices in the design of specialised and spoken corpora. The next section recounts the construction process through reference to the requirements above in combination with in-text examples.

Building the DAIS-C

Design statement

Sinclair's (2005) and Adolphs and Knight (2010, p.39) first recommendation about corpus building raises important questions about where language disturbance in schizophrenia sits in relation to corpus linguistic theory:

> *The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.*

Individuals with schizophrenia represent a speech community. They also experience a heterogeneous set of symptoms that affect linguistic production and comprehension (McKenna, 2007; McKenna and Oh, 2005). Some such symptoms correspond reliably to linguistic manifestations. The act of observing schizophrenia symptoms results in embedded assumptions about the linguistic potential of an individual's speech. Someone with pronounced negative symptoms has a good chance of showing poverty of speech, for example (Andreasen, 1982; Fervaha et al., 2016). I should build a corpus based not on these language features but instead on contents that reflect the 'communicative function in the community in which they arise' (Adolphs and Knight, 2010, p.39). The problem in this case, though, is that the contents that reflect those communicative functions also happen to imply specific language features. The present paper recounts how this was managed at the design and construction levels, in accordance with Sinclair's (2005) and Adolphs and Knight (2010, p.39) seventh recommendation:

> *The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.*

Ethics

Ethical review

A favourable ethical opinion was granted by the Health Research Authority (HRA; IRAS ID: 225295) following review by The South West - Plymouth and Cornwall REC on 3 July 2018. The study was also reviewed and approved by the Manchester Metropolitan University's Research Ethics and Governance team (EthOS ID: 5342) on 4 December 2018.

Informed consent

Participants were asked to provide separate statements of consent for audio recording and data archival/distribution (as per Leech et al. 2014 and Thompson 2005). Consent was also sought for GP notification, as the General Medical Council (GMC) recommends notifying participants' GPs, with their consent, of their involvement in research (GMC, 2013) regardless of their group allocation. Neither GP notification nor consent to archival were conditions of participation. The referring psychiatrists handled this process for clinical participants unless this was deemed unnecessary by the participant and/or their treating clinician. Comparison participants who consented to this were advised to share the latest participant information sheet (IS; v.3.2. 13 October 2020) with their GPs. All participants signed the latest approved version of the informed consent form (ICF; v.2.2., 13 October 2020) and medical declaration (MD; v.1.0, 12 February 2020) after reviewing the IS for a second time and raising any questions they might have had with the interviewer. IS and MD documents were distinct for each group (see Section 2.4.2).

Inclusion and exclusion criteria

Clinical group. Collaborating clinicians were asked to identify potentially eligible clinical participants who met the following inclusion criteria, from the latest approved study protocol (v.3.2., 13 October 2020):

- A formal, historic diagnosis of schizophrenia.
- Prescription of, and compliance with, antipsychotic medication (identified by referral and/or self-report).
- Deemed to hold capacity, and suitable, *via* SCA (structured capacity assessment).
- Referred by principal investigators and/or local collaborators.

Eligible potential participants were not approached or were withdrawn if they met the following exclusion criteria, from the latest approved study protocol (v.3.2., 13 October 2020):

- Comorbid neuropathology external to the scope of the research question—e.g. traumatic brain injury (identified by both the self-declaration and the CLQT—Cognitive Linguistic Quick Test Plus).
- Deemed unsuitable following SCA.
- Part one participants belonging to groups A and B who, due to a change of circumstances, no longer meet the relevant inclusion criteria and/or have since met the relevant exclusion criteria.

Comparison group. Comparison participants self-referred in response to public advertisement and needed to meet the following inclusion criteria, from the latest approved study protocol (v.3.2., 13 October 2020):

- No formal, historic diagnosis of schizophrenia.
- Deemed to hold capacity, and suitable, *via* SCA.

Eligible potential participants were withdrawn if they met the following exclusion criteria, from the latest approved study protocol (v.3.2., 13 October 2020):

- Comorbid neuropathology external to the scope of the research question—e.g. traumatic brain injury (identified by both the self-declaration and the CLQT—Cognitive Linguistic Quick Test Plus).
- Deemed unsuitable following SCA.
- Historic and/or current prescription of antipsychotic, antidepressant, and/or mood-stablising medication (identified by self-declaration).

These criteria are consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) fourth recommendation*:*

> *Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.*

The aim of the DAIS-C is to create a small, specialised, spoken language corpus that permits comparison of groups (clinical and comparison) on the basis of homogenous factors (such as interview question, mode of administration, etc.). This is consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) ninth recommendation:

> *A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.*

Here, the exclusion criteria offer protection against 'rogue texts'. Sinclair and Wynne (2004) define these as texts that stand out as unrepresentative of the variety in question. The homogenous distinctions between individuals with schizophrenia and nonpsychiatric comparison speakers are important for the creation of the DAIS-C and its distinct sub-corpora. Within these sub-corpora, it is important that samples taken from both groups of speakers are as free from competing clinical noise as possible. This is particularly important because, in the case of comorbid neuropathology, noise and signal are difficult to distinguish. The suggestion that FTD may represent a form of dysphasia, for example, is neither a conceptually nor linguistically light one.

Sampling

Clinical group

Collaborating clinicians applied a purposive, maximum variation sampling approach to the eligible participant pool. Clinicians did not select participants based on predefined linguistic criteria but rather an attempt to represent the range of symptom heterogeneity expressed in schizophrenia populations as a whole. This is consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) first and second recommendations:

> *The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.*

> *Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.*

It is reasonable to expect clinicians' familiarity with a participant's linguistic style to be problematic for the above. This was addressed by using an unstructured interviewing approach and blinding clinicians to the interview questions and their order of assignment. Clinicians were nevertheless aware of the broad study aims as listed on the IS and the general direction of interviews as described in the study protocol:

> *From the IS (v.3.2., 13 October 2020)*
>
> *'What is this study about?*
>
> *This study is about creative language and schizophrenia. I want to see if there is a relationship between creativity and the speech changes that can sometimes occur with schizophrenia. The findings from this study could advance our understanding of these speech changes and may prompt clinicians to think about language differently.*
>
> *What do you mean by creative language?*
>
> *By creative language, we mean the language of creative writers. Poets and novelists often break the 'rules' of language to achieve their effects: to inspire unique feelings, ideas, and perspectives.'*
>
> *From the protocol (v.3.2., 13 October 2020)*
>
> *'Participants will be asked to speak freely on the topic of their participation experiences and/or any other uncontested topics.'*

Reasonable attempts were made to ensure that clinicians could exercise clinical judgement about participant suitability and apply a maximum variation sampling approach that did not introduce significant linguistic bias in the form of their familiarity with potential participants' linguistic styles. This familiarity is closer to the definition of external (rather than internal) corpus construction criteria, as it is difficult to anticipate a participant's level and style of engagement when the line of questioning is not known: 'In general, external criteria can be determined without reading the text in question, thereby ensuring that no linguistic judgements are being made' (Atkins et al., 1992, p.8). Clinicians were surprised by the extent of variation in subject matter on viewing the transcripts, which suggests that the blinding was successful.

It is also important to point out that the interactional context on which the clinicians' familiarity is based differs substantially from that of the interviews and the corpus. These were informal conversations that performed no clinical or therapeutic function. This may have further helped to separate selection and corpus construction factors.

Another important factor guiding participant selection with reference to corpus construction is that clinicians are trained in the assessment of language pathology in a manner that differs from a detailed linguistic analysis. This also provides some protections against the (hypothetical but reasonable) view that clinician knowledge of speaker style may negatively impact the design.

Sampling a clinical population whose symptoms affect linguistic production and comprehension is challenging for corpus designers. Schizophrenia symptoms that affect language ability are arguably external criteria, yet their correspondence to specific linguistic manifestations makes it difficult to construct a corpus for this population that completely avoids building on internal criteria. The fact that symptoms correlate with certain manifestations, however, does not guarantee prediction of what language a corpus involving those symptoms will contain.

Comparison group

Self-selection sampling was used for comparison participants. Potential comparison participants responded to public advertisement. Their role in the corpus design process is much simpler. Self-selection sampling entirely avoids the problem of a corpus builder making linguistic judgements about speakers in this cohort.

Their lack of homogenous membership within a specific linguistic community is also beneficial as a point of contrast. This speaks to Sinclair's (2005) and Adolphs and Knight (2010, p.39) third recommendation:

> *Only those components of corpora which have been designed to be independently contrastive should be contrasted.*

The ability to compare speech in a clinical subgroup against that of a comparison subgroup was an integral aspect of the design. Early plans included no collection of comparison interview data due to the availability of the BNC. A general reference corpus does not, however, offer an increase in the homogenous components within the corpus. This can be achieved by including a comparison cohort and is consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) ninth recommendation:

> *A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.*

Relying on the BNC for comparison would have effectively produced a sub-corpus of 'rogue texts', because the DAIS-C's interactional contexts are not reflected in the BNC. It is preferable in the context of a specialised corpus to compare clinical and comparison speech drawn from the same interactional contexts. This leverages the main strength of specialised corpora.

Interviewing

FTD can be elicited in informal spoken contexts, particularly when the discussion focuses on open, abstract topics (Marengo et al., 1986). These interviews tend to be administered by clinicians in a test or clinical interactional context. Less work has looked at how individuals with schizophrenia converse in informal conversation on abstract topics.

Interviews were unstructured and involved three initiating questions. Only one such question was asked per participant, and two of the questions were randomised across participants. It was not possible to randomise the third question because it was reserved for participants who had also completed a psycholinguistic task. It was only used with one clinical speaker, and the referring clinician did not know in advance which of the referred participants would complete all measures or only the interview. All of the questions were about creative uses of language and did not broach clinical topics unless they were raised by the participants. These topics were only explored with participants' consent.

Three questions were developed to allow for representation of three concepts important to the research question.

The first is about whether creativity is defined narrowly or broadly. The psycholinguistic experiment described in Delgaram-Nejad et al. (2022) offers a narrow, experiential definition. In that experiment, participants may make creative choices but only under restrictive linguistic conditions. This question is important for eliciting information on broad concepts confined to a limited experiential frame. The opposite of this involves inviting participants to define linguistic creativity themselves. This provides more opportunity for digression and abstraction but can be intimidating for respondents. Varying these question types allowed for data capture at both extremes.

The second concept was the role of an open or closed initiating question. Open questions invite a range of responses, whereas closed offer less (usually affirmative or negative). As an initiating question, closed questions allow for the quick categorisation of participants' viewpoints before the reasoning is unpacked with a subsequent open question. It was important to capture data on both question types in the initiation (or cue) position for two reasons: (1) because wh-questions and closed questions exert different effects in interview contexts (Waterman et al., 2001); (2) because some individuals with schizophrenia perform poorly on tests of social cognition and open questions place greater demands on those (and broader cognitive) resources; question type studies in childhood-onset schizophrenia suggest difficulty with wh-questions independent of cognitive functioning and/or the presence of FTD (Abu-Akel et al., 2000). A corpus including data reflective of both types provides helpful information on interviewing styles as they relate to schizophrenia cohorts, response formulation and structure by genre, and more.

The third concept was about whether emphasis fell on language or creativity. One question framed creativity as an action that could involve language, whereas another framed language as a tool that could be exploited for creative purposes. This ensured that responses within the corpus were generated from a variety of conceptual prompts.

These variations in question type aimed to be consistent with the idea that samples should be taken from a range of typical situations (Almut, 2010). Descriptions of interview behaviour tend to make up a smaller part of corpus design (as in Pedraza 2019).

The questions are as follows:

[1] *'How was the experiment?'*

This question was used only with participants who had completed the psycholinguistic task described in Delgaram-Nejad et al. (2022). It was intended to gather data for the 'open initiating' and 'narrow creativity context' genres.

[2] *'Do you feel like you do creative things with language?'*

This question was randomised across all participants who did not take part in the psycholinguistic task. It was intended to gather data for the 'closed initiating' and 'broad creativity context' genres. Emphasis was placed on affect ('feel' - to prompt abstract reasoning) and creativity as an action that could involve language ('do creative things with').

[3] *'Do you feel like you use language creatively?'*

This question was also randomised across all participants who did not take part in the psycholinguistic task. It was intended to gather data for the 'closed initiating' and 'broad creativity context' genres. As above, emphasis was placed on affect ('feel' - to prompt abstract reasoning) and, in this case, language as a tool that could be exploited for creative purposes ('*use* language creatively').

These decisions represented attempts to observe Sinclair's (2005) and Adolphs and Knight (2010, p.39) eighth recommendation:

> *The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components. Any comparison of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.*

The interviewer used an unstructured approach with the chief goal of maximising ecological validity. The interviewer spent the interaction processing participants' responses to one of the initiating questions and developing follow-up questions online (i.e. in real time). This approach was about communicating interactional parity, as both interviewer *and* interviewee had to formulate their contributions in real time. The interviewer would ask for clarification of specific concepts and elaboration on certain terms, focusing on points of metalinguistic awareness. The interviewer reintroduced creativity as a topic only if the participant had deviated significantly *and* reached the point where they could no longer advance the conversation themselves. The interviews were concluded when participants indicated that they had said all that they wished to. They were advised at the start of the conversation that they could do this at any time, and the interviewer checked participant views on this at various points throughout the interview. The interviewer signalled this point in the interaction with a closed (and closing) question: 'is there anything that you'd like to talk about that we haven't talked about?' This is consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) sixth recommendation:

> *Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.*

Many participants sought a definition of linguistic creativity, even though (a generic and somewhat nonspecific) one was provided on the IS. The interviewer provided their own opinions on this and other topics when asked, again to sustain ecological validity and communicate interactional parity. The interviewer would even offer alternative points of view. These were introduced because naturalistic interaction regularly requires the navigation of difference, something often missing from traditional qualitative interviewing paradigms. The interviewer never insisted upon their views, though, and events where this was situationally appropriate were rare.

Recording

Audio quality, interview duration, recording date, and recording time were documented (as per Strassel and Cole 2006). This allows for calculations of words per audio minute and other analyses.

Information about the devices used by the interviewer and participant and interview recording arrangements was also logged to contextualise the audio quality tables. Some participants were interviewed *via* telephone, with the speakerphone function activated, which was then recorded using a desktop condenser microphone. This degrades the final audio signal because the speech data is filtered at several points. This had transcription implications that are discussed later (see Section 2.5.) and was a significant factor in the need to exclude <26AR12>'s data.

Demographics

Age, sex, gender, and education

Data on biological sex (female or male) and gender identity (woman, man, or a specified alternative) were recorded because both influence outcomes in clinical research (Clayton and Tannenbaum, 2016). Data on age range and education level is missing for some clinical participants due to errors in data collection, although some education information has been recovered as it is referenced in the transcripts. Education was recorded as positioning in relation to the UK National Qualifications Framework (NQF).

Setting and geography

A design benefit of offering multiple participation routes (in-person, remote, all measures, interview only) was that it provided situational variety (Almut, 2010). Genres and registers can be well represented with samples spanning the five to ten range (Biber, 1990).

Transcription

Conventions

Transcription conventions were developed by modifying those present in the BNC User Manual and Reference Guide (v.1.1., Lancaster University, 2014). The original BNC formatting and approaches were retained wherever possible. When adaptations were necessary, they were designed to work with existing BNC conventions.

This part of the design stage was about identifying textual, prosodic, gestural, and environmental elements (as per Adolphs and Knight 2010) useful to the present study of linguistic creativity and schizophrenia. Attention was given to features that might also benefit other researchers (especially where they coincide with the aims of the current study). The general aim was a lightweight set of broadly useful conventions that capture those environmental properties that contribute to the multimodality of spoken language (Strassel and Cole, 2006). The selection of transcription elements therefore focused on those of potential relevance to spoken interactions overall, clinical cohorts and creativity, *and* the broadest levels of linguistic analysis: phonology, morphology, syntax, semantics, and pragmatics.

Table 1, below, displays the conventions used in the DAIS-C.

Most conventions follow the Extensible Markup Language (XML) format, having an opening and closing tag. This decision was made to improve the end-user experience, especially within corpus analysis software. This approach also allows those interested in more granular analyses to situate these features at any point in the raw text, as in the following example:

> *<Lh>al</Lh>right then*

This approach was not required for the present study, but the text has been prepared such that others can adopt this approach if they choose.

XML tags are also easy to extract en-masse, allowing for rapid and precise token/word counts of both features of interest and of the raw text. This approach is consistent with Sinclair's (2005) and Adolphs and Knight (2010, p.39) fifth recommendation:

> *Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.*

Speaker labels are composed of participants' unique identifiers, which were developed using the study debrief sheet (DS; v.1.1, 13 June 2018).

Laughter, coughing, sighing, and sniffing were included because they provide valuable paralinguistic information about participant status, potential emotional state, and so on. Lateral clicks were recorded for both their paralinguistic value and for their potential relevance to EPSEs and clozapine (Li et al., 2009). Inaudible speech was coded to provide a qualitative account of missing speech. It was not possible to discriminate the number of inaudible utterances in all cases, and so the convention is catch-all for the word and phrase levels. Miscellaneous noise, given its variability, was covered with a single code that allowed for transcriber comments. Specifics about the nature of the noise (e.g. whether it was a human voice, a motor vehicle, or music in the distance) was added within this layout. Anonymised information and missing data are treated separately for record-keeping purposes.

Anonymisation was carried out at the transcription stage. All personally identifiable information (PII), such as names and e-mail addresses was redacted and processed in accordance with the latest approved study protocol (v.3.2. 13 October 2020). Data that was not PII but may still have been identifying in some circumstances was also removed (as per Du Bois. 1992). Examples include reference to frequented locations and landmarks, sites of previous hospital admission, and anecdotes about other people. There are instances where the identify of public figures can be inferred despite anonymisation, such as the following:

> *but princess <An> </An> and I met <An> </An> he came to er the hospital in <An> </An> princess <An> </An> and like the patients were very upset because then you got celebrities coming in and just taking the piss out of the patients y I mean you know and then you have people like er <An> </An> lady <An> </An> I do like lady <An> </An> she's the one person that I like in the royal family yeah*

This is not problematic in itself as these figures are widely known. Thought was given, though, to whether this speaker's reference to meeting said figures while an inpatient could be identifying. The ultimate determination was that it could not, as none of the details, when anonymised and combined, indicated without doubt any one hospital or occasion.

Hyphenation was avoided for compounds but retained for orthographic number (e.g. 'fifty-five'), as were pauses and sentence boundaries. It became clear early in the transcription process that any attempt to delineate sentence units (as in s-units in the BNC) in the more disorganised examples required considerable time and effort. Given that doing so would not be of great benefit to the research question, it has not been done in this version. It would be possible to introduce them later, however.

Pauses were annotated in early transcription attempts. This greatly slowed transcription, which was already taking some time. There was also much agonising over the value that timed pauses would offer other researchers, their potential relevance to speech disorganisation in schizophrenia (given what they may reveal about executive function), and the work involved in their inclusion. Pauses were ultimately dropped, and the audio files were destroyed on the production of a final transcript. The knowledge that the files could not be retained for further transcription also shaped the approach taken here. A great deal of data had to be discarded, such as detailed prosodic information of potential interest to speech and language therapists and phoneticians. The desire to transcribe with ever-increasing precision (Cook, 1990; McCarthy 1998; Carter, 2015; Halliday, 2004) was particularly apparent at this stage. To ensure that not all prosodic information was lost, an economical (and unusual) form of (what might be called onomatopoeic) transcription was employed (shown below):

> *<21AN11> because they see something good but they don't like it it it upsets them like they have a problem with me listening y I mean if I was a lis er f if I'm a very good listener y I mean I I I listen in reality I listen crystal clear it's mm I may not think the other way the way other people think but I think the way I think </21AN11>*

The phonetic properties of the participant's speech are represented in examples like 'y I mean', and truncated words are presented as they sounded on the recording 'lis er f if'. This form of representation was chosen because the DAIS-C is not a written corpus, dysfluencies and their articulatory properties are potentially relevant to the research question and certainly relevant to the language community and interactional context under study, and the method (although non-standard) allows for the detailed representation of phonological information without the use of intensive phonemic or phonetic annotation. It also avoids the problem of estimating the intended word in the case of truncation, which was often not possible with any confidence. In the rare cases where this form of representation conflicted with standard orthography, for example where 'well' truncated to 'we' would lead to confusion with the pronoun, an alternative that still conveyed the main concept was used: 'w'. This is a good example of how transcription is indeed highly theoretical (Ochs, 1979; Edwards, 1993; Thompson 2005). This approach attempts to tread the difficult line between validity and reading ease (Graddol et al., 1994).

Table 1
DAIS-C transcription conventions, modified from the BNC 2014.

Tags	Description
<XXXXXX> </XXXXXX>	speaker label/ID
<INT> </INT>	interviewer speech
<FAM> </FAM>	family members
<DOC> </DOC>	clinicians
<Lh> </Lh>	laughing
<Ch> </Ch>	coughing
<Sh> </Sh>	sighing
<Sn> </Sn>	sniffing
<Cl> </Cl>	lateral clicking
<InAu> </InAu>	inaudible speech
<Noi=description> </Noi>	miscellaneous noise
<An> </An>	anonymised data
<Mis> </Mis>	missing data
mm	voiced pause
mhm	voiced pause, affirmative
er	filler sound, as in 'her'
erm	filler sound, as in 'term'
ah	filler sound, as in 'car'
oh	filler sound, as in 'toe'
ay	filler sound, as in 'stay'
w, wh, I, la	truncated words
cos, wanna, gotta	standardised contractions
pleisure, P L E I S U R E	words spelled aloud

Software

EasyTranskript, a free-to-use transcription environment, was used to process audio files. The software was chosen because it allows for the quick production of timestamps. These are provided as a separate file group within the corpus file structure.

Storage

The corpus data are presented variously across a range of file types and formats, to counter the problem of 'one-offness' common to the development of specialised corpora (Almut, 2010).

Interactional files contain both interviewer and participant dialogue, presented in a running sequence as shown below:

<INT> wow </INT>

<03EB14> so but that it it that is about it really that is about it so I write a poem <InAu> </InAu> it's the first time in a long long time but it is mostly songwriting that I get creative with so </03EB14>

<INT> what's the the thing you like about the songwriting more than the poetry </INT>

Timestamped files mirror the interactional files but contain only the timestamps associated with each speaker's turn, as shown below:

<03EB14> #00:06:11-5# </03EB14>

<INT> #00:06:08-4# </INT>

<03EB14> #00:06:34-6# </03EB14>

Disruptions in the chronological order, as above, can be used to infer overlaps. This is because timestamp markers correspond to the end of each speaker's utterance, irrespective of the order of turns.

'Speaker Only_XML' files contain only the participant's turns and all XML annotation, as shown below:

<03EB14> yeah erm I I sorry </03EB14>

<03EB14> erm I I use erm what I do is erm I will I w I will build like a a single PNG like what what w one moment one d is one way of doing it is to create many individual

'Speaker Only_Raw' files contain only the participant's turns, with all but the plain text removed as shown below:

erm yeah

is that what

sorry

erm

I erm could be may erm that erm erm if I if I hadn't hadn't had all this erm like say like bad stuff in

These decisions focus mainly on Sinclair's (2005) and Adolphs and Knight (2010, p.39) fifth recommendation:

> *Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.*

DAIS-C characteristics

Corpus characteristics

The tables below show the total number of tokens, audio hours, and audio minutes across the DAIS-C as a whole and also by sub-corpora. It is important to note that tokens account for speakers only, whereas audio hours and minutes account for the interaction as a whole (speaker and interviewer). Tables 2–4 display this information.

Speaker characteristics

Overrepresentation

The corpus is characterised by two forms of overrepresentation at the group level. The first is the distribution of females and males and the second is the distribution of interview contexts.

Tables 5–7 show this in cross tabular format.

Education is arguably a third source of overrepresentation, as comparison participants range L6 to L8 on the NQF. A full comparison against the clinical cohort is not possible due to insufficient data, though there is data suggesting clinical representation of L3, L6, and L7.

The above are clearly problematic for Sinclair's (2005) and Adolphs and Knight (2010, p.39) ninth recommendation:

> *A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.*

The design included plans to recruit an equal number of males and females and to spread the questions and contexts evenly across groups. Achievement of demographic representation was largely dependent on illness severity and suitability, availability, and interest in or inclination toward taking part in the study. Variations in participation route arose mainly in response to the situational and operational challenges presented by the pandemic, and question type was tied to the experimental approaches, particularly the block randomisation processes, described in Delgaram-Nejad et al. (2022).

Schizophrenia and FTD affect language acutely, chronically, and markedly (McKenna, 2007; McKenna and Oh, 2005). Sociolinguistic influences relating to sex and education follow a different pattern and course. Context has a more acute effect, but sociolinguistic influence and language pathology are sufficiently distinct to allow for the detection of schizophrenia-specific effects within the DAIS-C. Their sociolinguistic and contextual dependencies will require a larger corpus (but perhaps not a reference corpus, as covered in Delgaram-Nejad et al. (2022) and further study, and corpus expansion can correct the demographic imbalances of this early version.

Corpus data

Tables 8 and 9 below show the number of tokens per speaker and their contribution to the total corpus and their respective sub-corpora as percentages.

Fig. 1 shows the distribution of token counts.

13 comparison speakers and 11 clinical speakers are within the 0–5000 token range. One comparison speaker and three clinical speakers are in the 5000–10,000 token range. One clinical speaker is an outlier, being within the 15,000–20,000 token range.

Fig. 2 below shows the distribution of audio minutes per speaker.

Eight comparison speakers and five clinical speakers are within the 0–20 min range. Four comparison speakers and six clinical speakers are within the 20–40 min range. Two comparison speakers and two clinical speakers are within the 40–60 min range. Two clinical speakers are outliers, one being within the 80–100 min range and another being within the 100–120 min range.

These values suggest some success in sampling according to Sinclair's (2005) and Adolphs and Knight (2010, p.39) sixth recommendation:

Table 2
Token count, audio duration, and audio minutes.

	Tokens	Audio hours	Audio minutes
DAIS-C	97,357	21:24:50	1284.8
DAIS-C-CL	58,444	15:47:28	947.5
DAIS-C-CO	33,025	05:37:22	337.4

Table 3
Token count, mean, and standard deviation.

	Tokens	Mean	SD
DAIS-C	97,357	3154	3011
DAIS-C-CL	58,444	3896	3700
DAIS-C-CO	33,025	2358	1864

Table 4
Audio minutes, mean, and standard deviation.

	Audio minutes	Mean	SD
DAIS-C	1284.8	29.21	22.31
DAIS-C-CL	947.5	34.97	27.12
DAIS-C-CO	337.4	23.04	14.19

Table 5
Sex.

		Female	Male	Total
Group	Clinical	3	12	15
	Comparison	11	3	14
Total		14	15	29

Table 6
Interview only.

		No	Yes	Total
Group	Clinical	1	14	15
	Comparison	10	4	14
Total		11	18	29

Table 7
Topic breadth.

		Broad	Narrow	Total
Group	Clinical	14	1	15
	Comparison	4	10	14
Total		18	11	29

Table 8
DAIS-C-CL: tokens, % of corpus, and % of sub-corpus.

Speaker ID	Tokens	% of corpus	% of sub-corpus
<03AR15>	3800.00	3.90	6.50
<03EB14>	15,473.00	15.89	26.47
<10EB14>	2345.00	2.41	4.01
<10EB15>	4274.00	4.39	7.31
<12AY15>	2040.00	2.10	3.49
<18UG09>	2344.00	2.41	4.01
<18UG10>	186.00	0.19	0.32
<18UG11>	2288.00	2.35	3.91
<18UG14>	5239.00	5.38	8.96
<18UG15>	1066.00	1.09	1.82
<21AN11>	6499.00	6.68	11.12
<21AP15>	4137.00	4.25	7.08
<21UN11>	1686.00	1.73	2.88
<22AP15>	1179.00	1.21	2.02
<26AR16>	5888.00	6.05	10.07

> *Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.*

One challenge associated with this recommendation is that samples of various sizes can skew relative contributions to the corpus, with some speakers constituting a much larger portion than others. It is worth reviewing each speaker's contributions on the level of their specific sub-corpora and that of the wider corpus.

Table 9
DAIS-C-CO: tokens, % of corpus, and % of sub-corpus.

Speaker ID	Tokens	% of corpus	% of sub-corpus
<02AR17>	819.00	0.84	2.48
<09AR14>	1693.00	1.74	5.13
<11AR18>	3371.00	3.46	10.21
<16OV11>	477.00	0.49	1.44
<16UN13>	4210.00	4.32	12.75
<16UN16>	4867.00	5.00	14.74
<17AR13>	1289.00	1.32	3.90
<19OV10>	1692.00	1.74	5.12
<23CT18>	1330.00	1.37	4.03
<23CT19>	939.00	0.96	2.84
<23EB14>	7031.00	7.22	21.29
<23OV14>	1842.00	1.89	5.58
<26CT11>	1210.00	1.24	3.66
<28CT11>	2255.00	2.32	6.83

Fig. 3 below shows each speaker's contribution to their relevant sub-corpus as a percentage.

14 comparison speakers and 13 clinical speakers are within the 0–10% range. One clinical speaker is in the 10–20% range. One clinical speaker is an outlier, being within the 20–30% range.

Fig. 4 below shows each speaker's contribution to the overall corpus as a percentage.

6 comparison speakers and 9 clinical speakers are within the 0–5% range. Four comparison speakers and three clinical speakers are within the 5–10% range. Three comparison speakers are in the 10–15% range. One clinical speaker is in the 15–20% range. One comparison speaker is in the 20–25% range.

The DAIS-C incorporates complete speech events, resulting in samples varying significantly in size. The rates of these variations are somewhat balanced across groups, but speaker overrepresentation is apparent at the sub-corpus and corpus levels. The result is a fair compromise between numerical uniformity and participant heterogeneity.

It is also important to review interviewer token data, as differences in interviewer behaviour are likely to influence participant behaviour.

Fig. 5 below shows the distribution of interviewer token counts.

Interviewer tokens range 0–1000 for four comparison and five clinical speakers. Interviewer tokens range 1000–2000 for seven comparison and eight clinical speakers. Interviewer tokens range 2000–3000 for two comparison and two clinical speakers. Interviewer token data is missing for one clinical participant: <23EB14>.

Fig. 6 below shows the distribution of speaker to interviewer token ratios.

Token count ratios (describing the number of speaker tokens for each interviewer token) are relatively balanced across groups. 13 (14 including the missing value for <23EB14>) comparison ratios and 12 clinical ratios fell within the 0–5 range. There are two clinical outliers in the 5–10 range. There was one clinical outlier in the 15–20 range.

Group characteristics

Overrepresentation

It is worth examining the speaker characteristics discussed in Section 3.2.2. In relation to the overrepresentation issues presented in Section 3.2.1.

Mean differences by sex. Fig. 7 below shows the mean token counts across groups by speaker sex.

Female comparison participants show a higher token mean than male comparison participants, with the means and dispersions in this group being somewhat similar. Female clinical participants show a

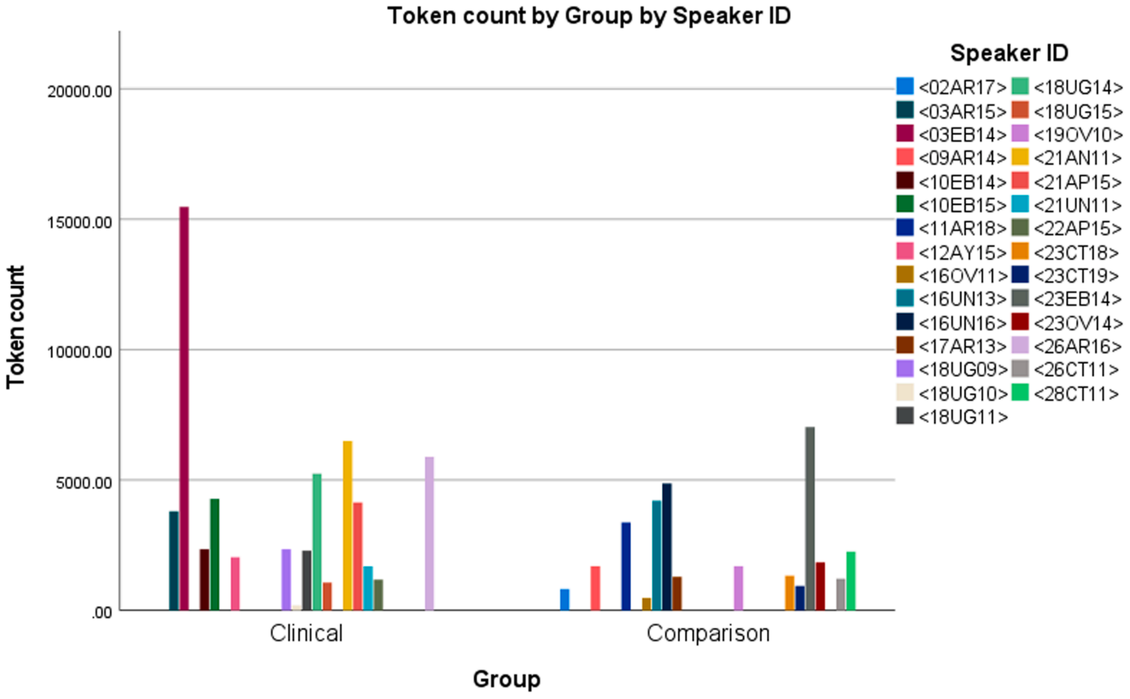


Fig. 1. Token count by group by speaker ID.

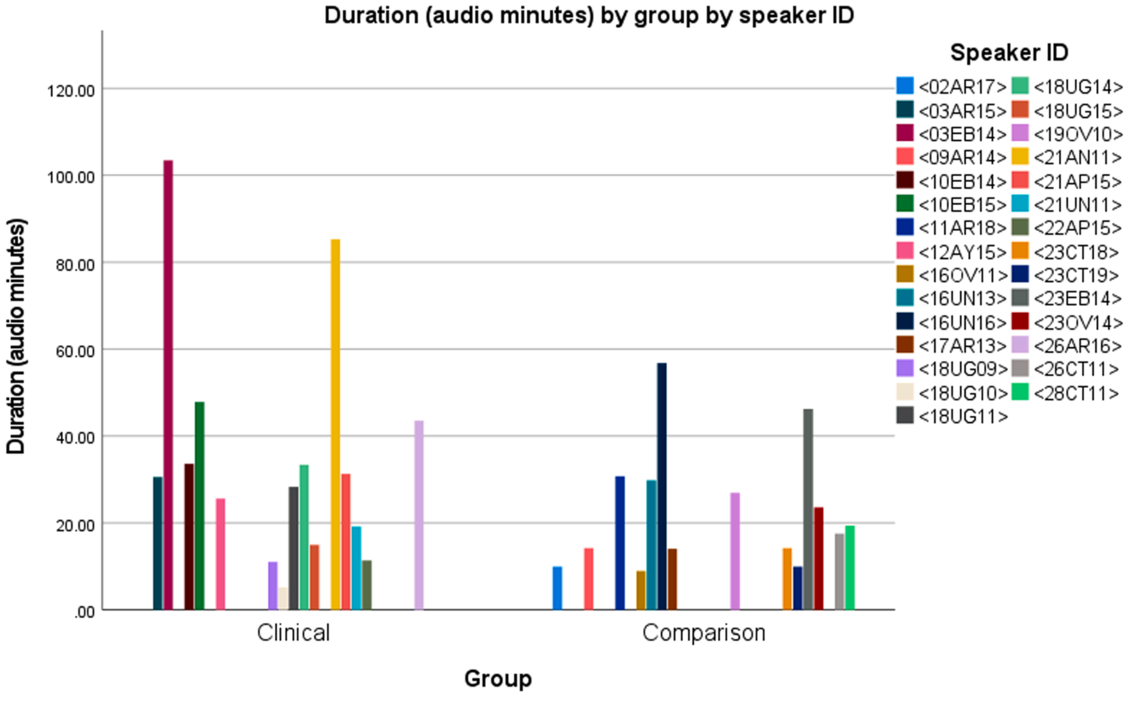


Fig. 2. Duration (audio minutes) by group by speaker ID.

higher token mean than male clinical participants, despite male overrepresentation in this cohort, alongside a smaller dispersion than the male clinical participants and comparison participants overall. Male clinical participants show a higher token mean than comparison females and males but with the overall largest dispersion across groups. Readers are reminded of the following: comparison females ($n = 11$), comparison males ($n = 3$), clinical females ($n = 3$), clinical males ($n = 12$).

Fig. 8 below shows duration in audio minutes by group by sex.

Female comparison participants show a higher duration mean than male comparison participants, with the comparison means and comparison dispersions being somewhat similar. Female clinical participants show a higher duration mean than male clinical participants, despite male overrepresentation in this cohort. And, again, female clinical participants display a smaller dispersion than the male clinical participants and comparison participants overall. Male clinical participants show a higher token mean than comparison females and males but with the overall largest dispersion. Readers are reminded of the following: comparison females ($n = 11$), comparison males ($n = 3$), clinical females

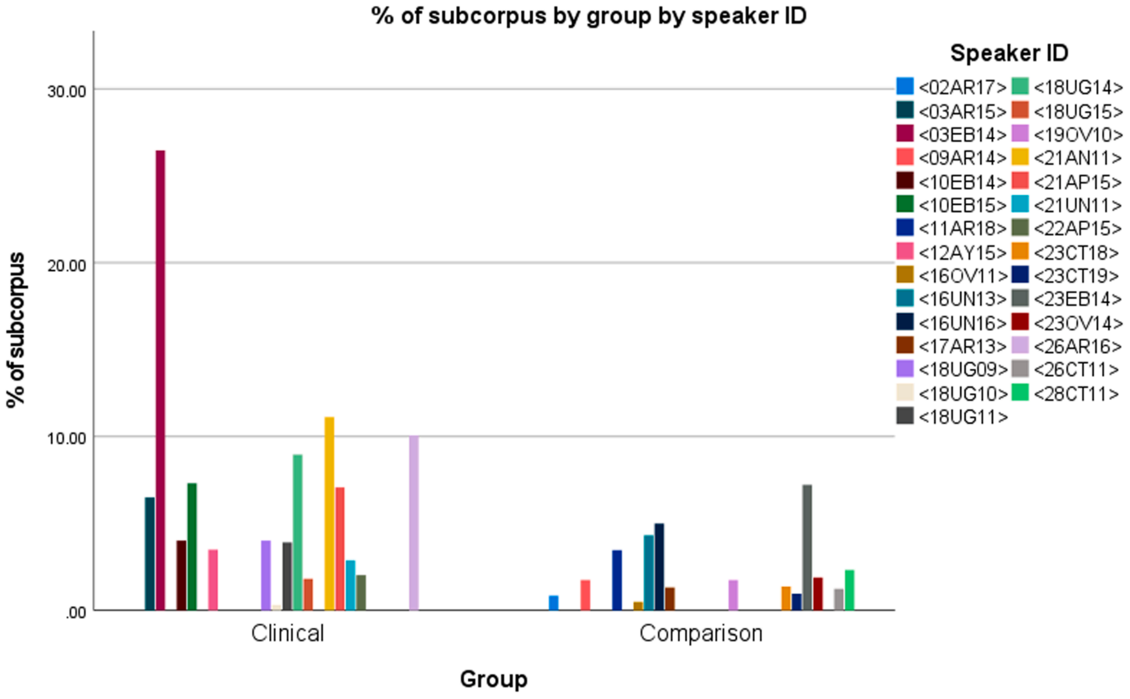


Fig. 3. % of subcorpus by group by speaker ID.

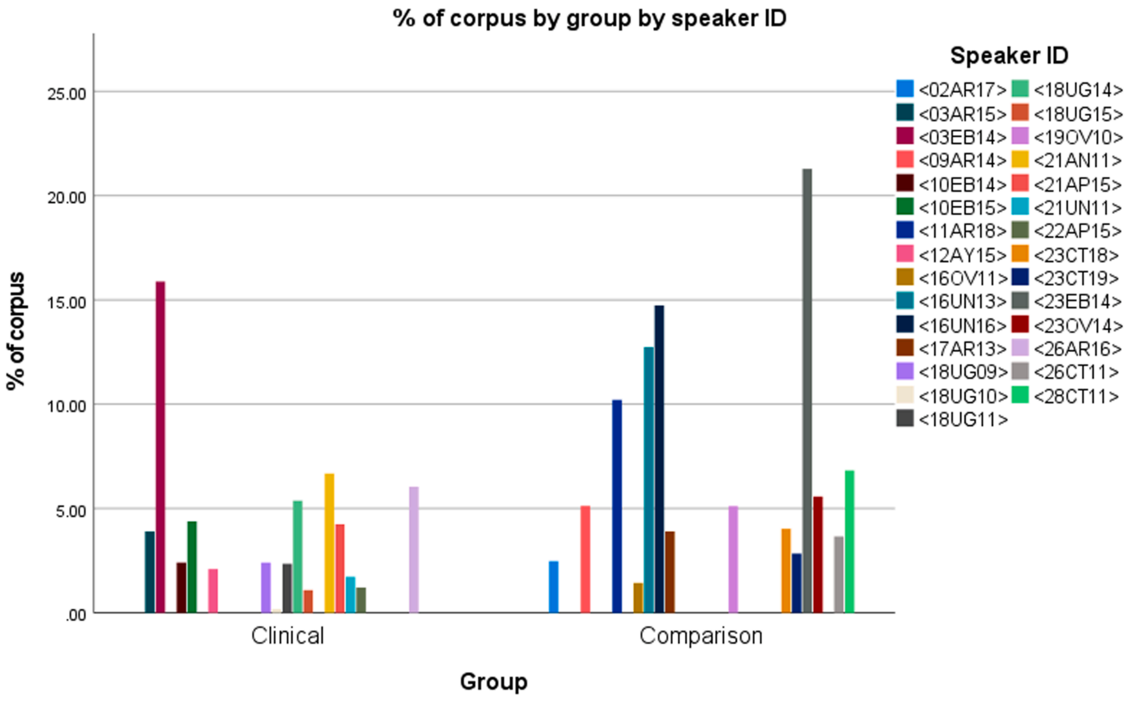


Fig. 4. % of corpus by group by speaker ID.

($n = 3$), clinical males ($n = 12$).

Fig. 9 below shows mean % of sub-corpus by group by sex.

Female comparison participants show a higher sub-corpus contribution than male comparison participants, with the comparison means and comparison dispersions being somewhat similar. Female clinical participants show a higher sub-corpus contribution than male clinical participants, despite male overrepresentation in this group. And, as above, there is a smaller dispersion in the female clinical subgroup than the male clinical participants and the comparison participants overall. Male clinical participants continue to show the largest dispersion. Readers are reminded of the following: comparison females ($n = 11$), comparison males ($n = 3$), clinical females ($n = 3$), clinical males ($n = 12$).

Fig. 10 below shows mean % of corpus by group by sex.

Female comparison participants show a higher corpus contribution than male comparison participants and clinical participants overall, with the comparison means and comparison dispersions being somewhat similar. Notably, comparison dispersions are larger than that seen

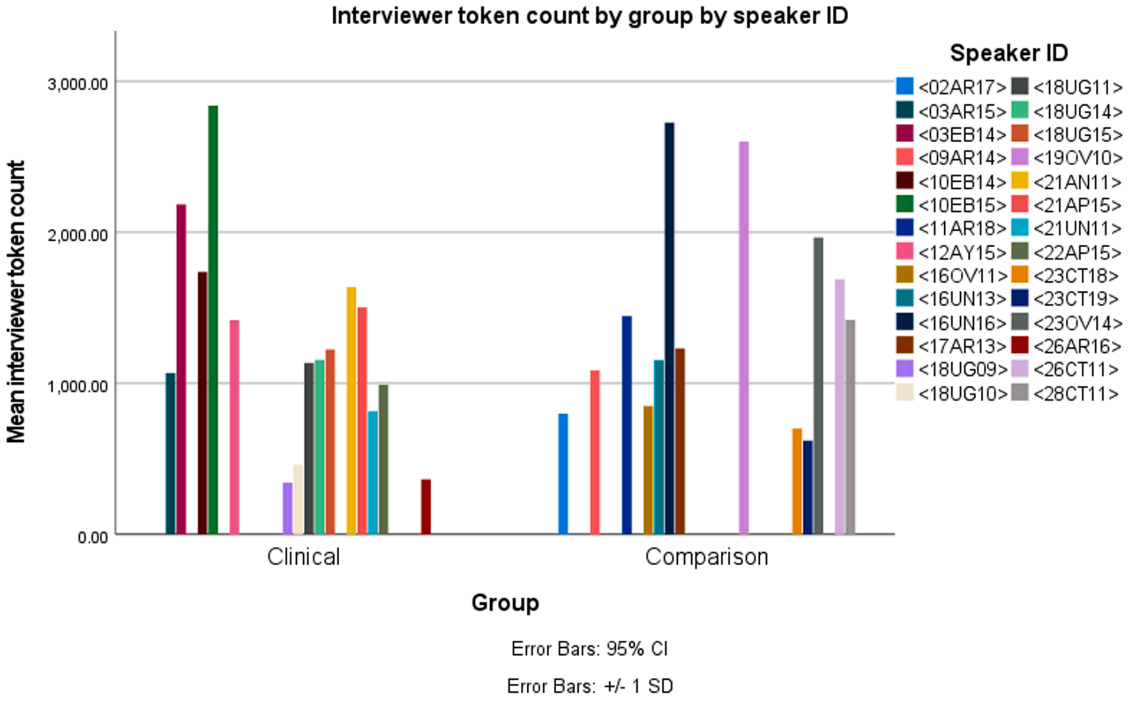


Fig. 5. Interview token count by group by speaker ID.

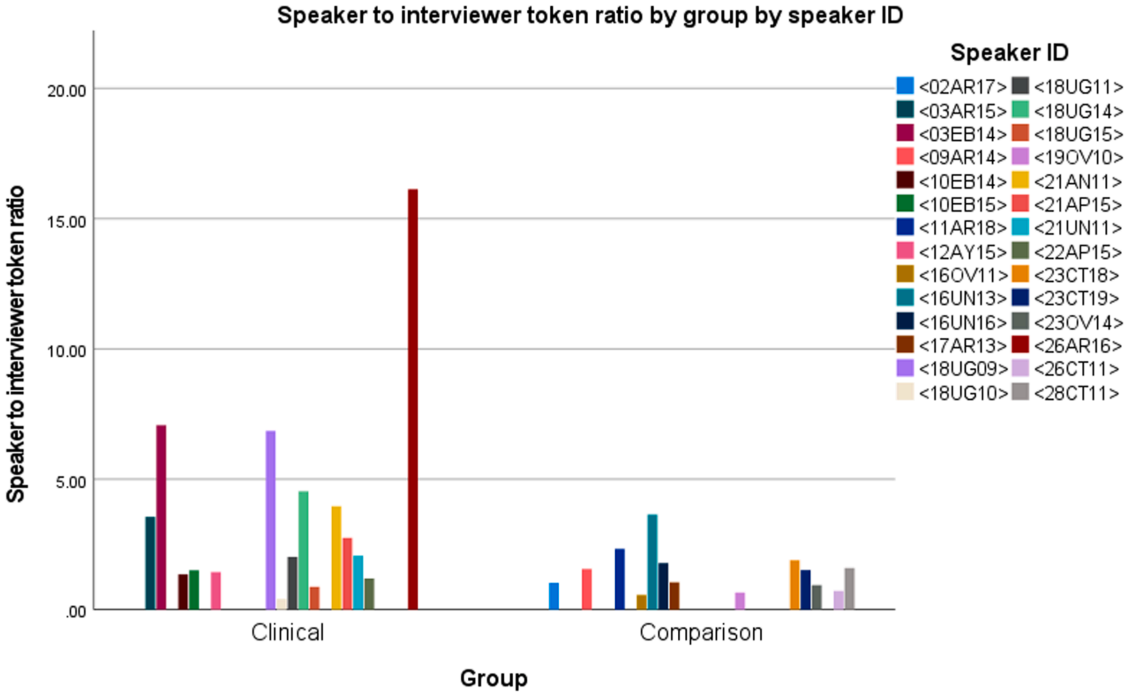


Fig. 6. Speaker to interviewer token ratio by group by speaker ID.

for clinical males. Female clinical participants show a higher corpus contribution than male clinical participants, despite male overrepresentation in this group. As above, female clinical participants display a smaller dispersion than male clinical participants and the comparison participants overall. Male clinical participants show the lowest corpus contribution, despite being the over representative demographic in the clinical group. Readers are reminded of the following: comparison females ($n = 11$), comparison males ($n = 3$), clinical females (n = 3), clinical males (n = 12).

In the DAIS-C, on average, female clinical participants produced more tokens, spoke for longer, contributed more to their respective subcorpus, and contributed more to the overall corpus than clinical males despite being demographically underrepresented at a ratio of 1:4. Female clinical participants also showed the least variance across all sex indices.

3.1.1.2. Mean differences by topic breadth (open versus closed: broad versus narrow). Fig. 11 below shows the mean token counts across

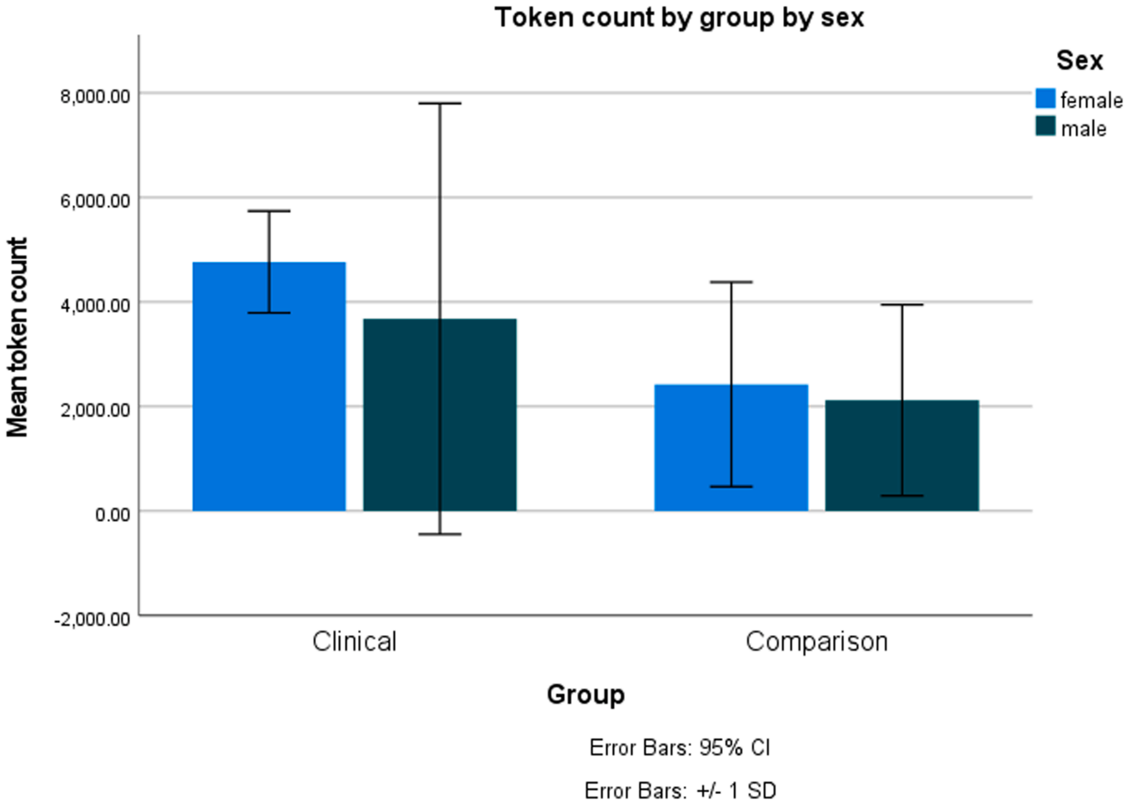


Fig. 7. Token count by group by sex.

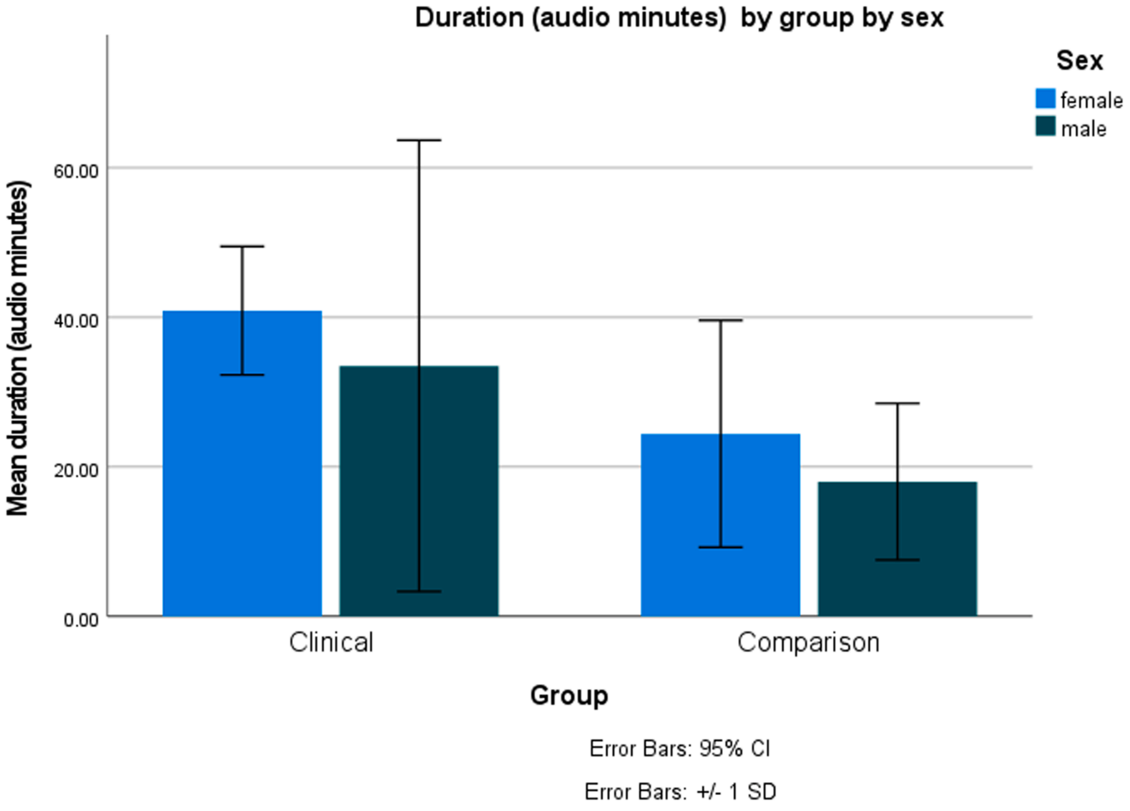


Fig. 8. Duration (audio minutes) by group by sex.

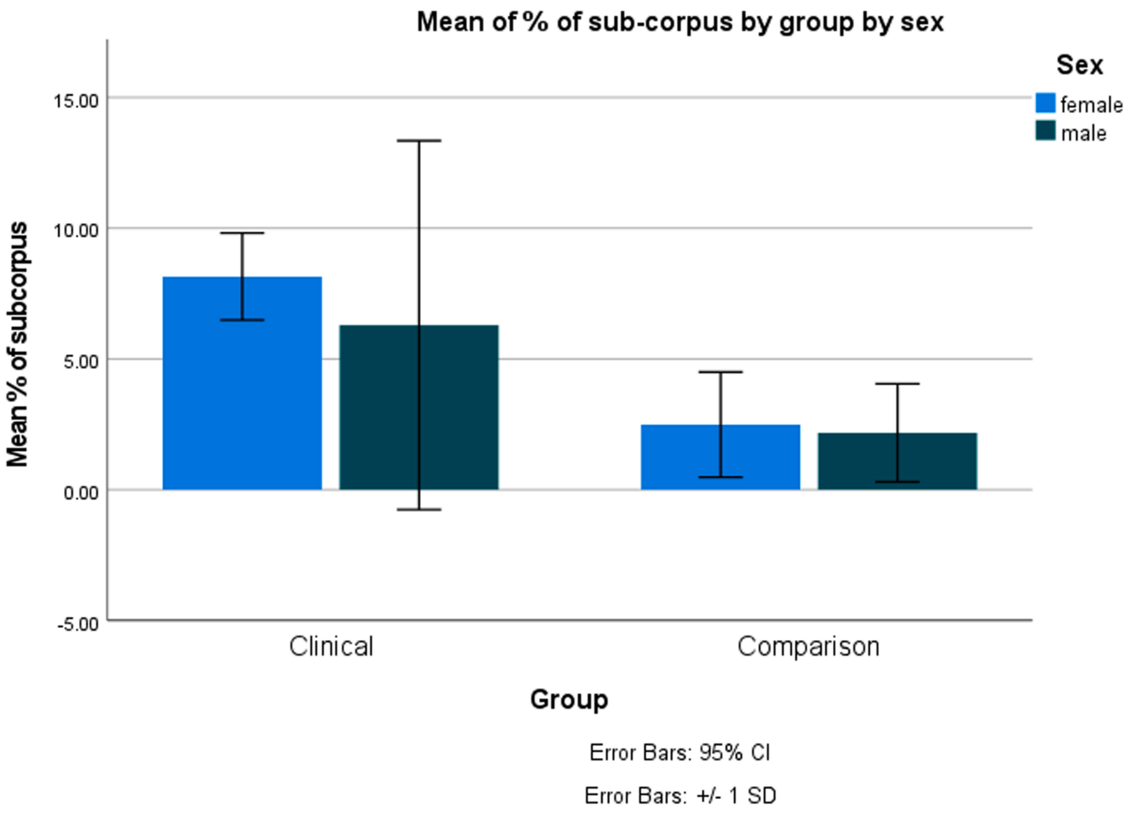


Fig. 9. Mean % of sub-corpus by group by sex.

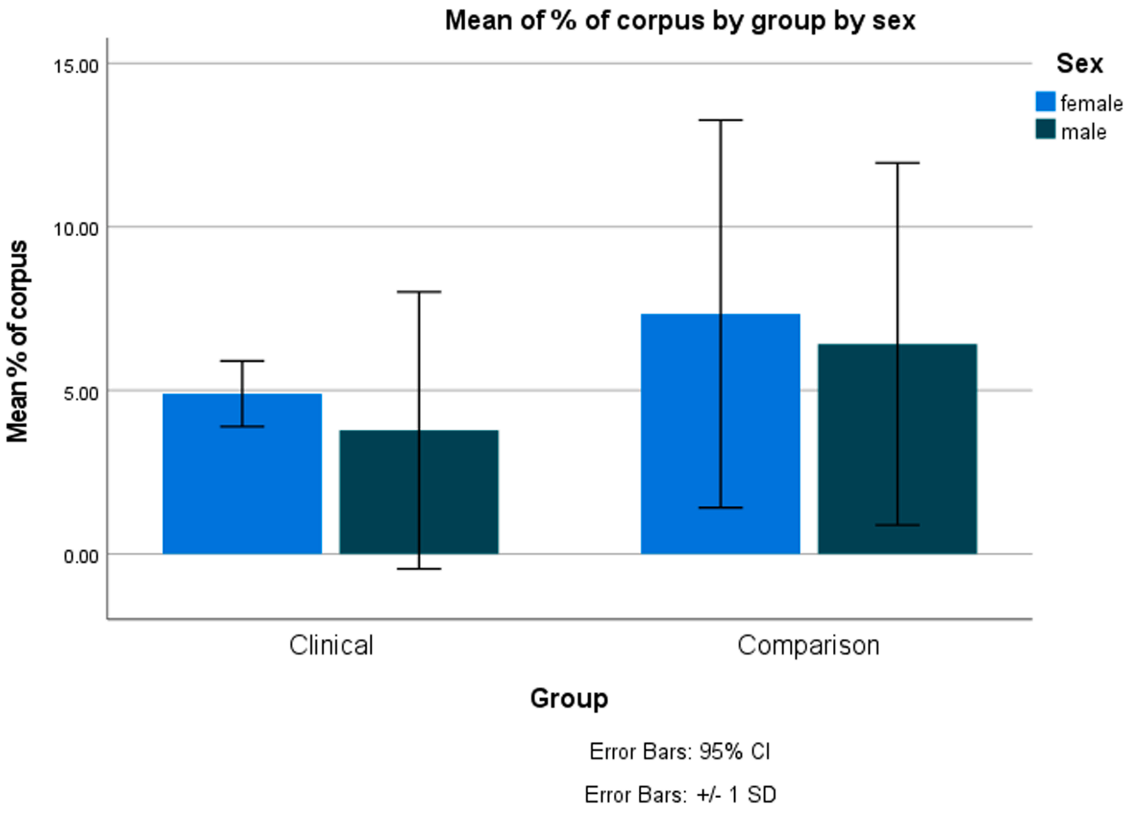


Fig. 10. Mean % of corpus by group by sex.

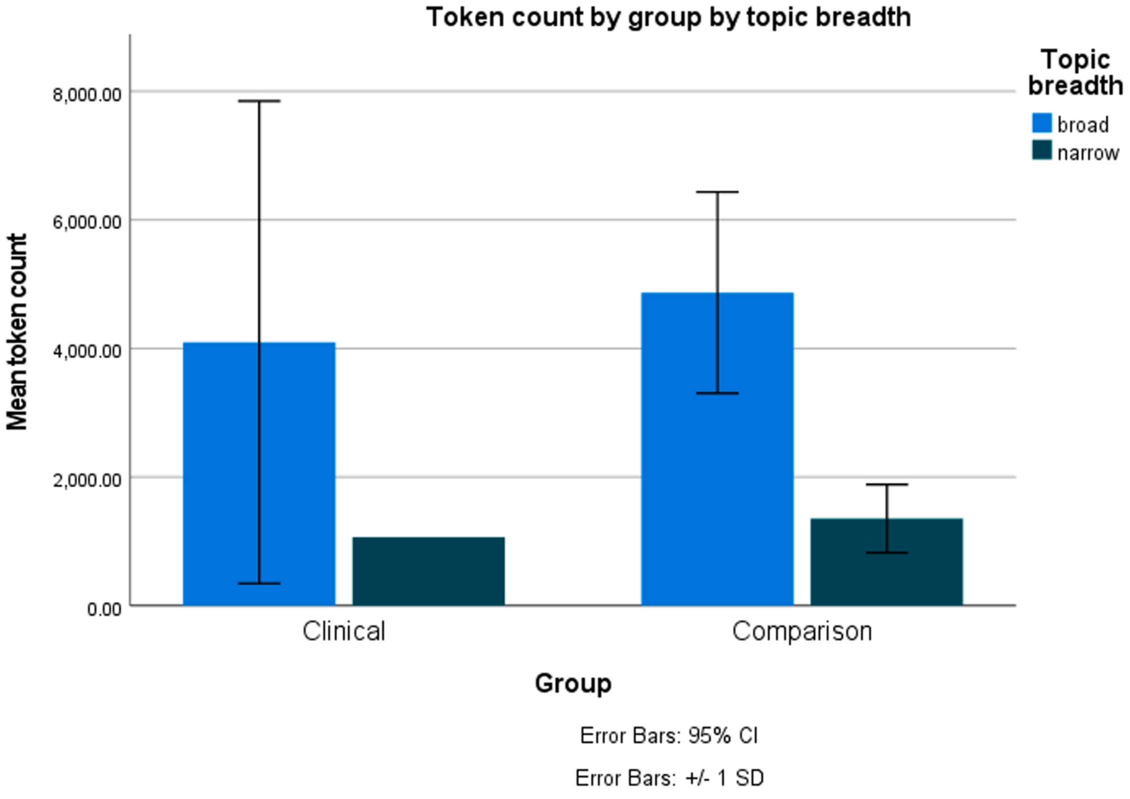


Fig. 11. Token count by group by topic breadth.

groups by topic breadth.

Comparison participants show a higher token mean on the broad topic than the narrow topic and the highest token mean across groups, despite overrepresentation of the narrow topic in the comparison group and overrepresentation of the broad topic in the clinical group. Comparison dispersions were closer to the comparison means. Clinical

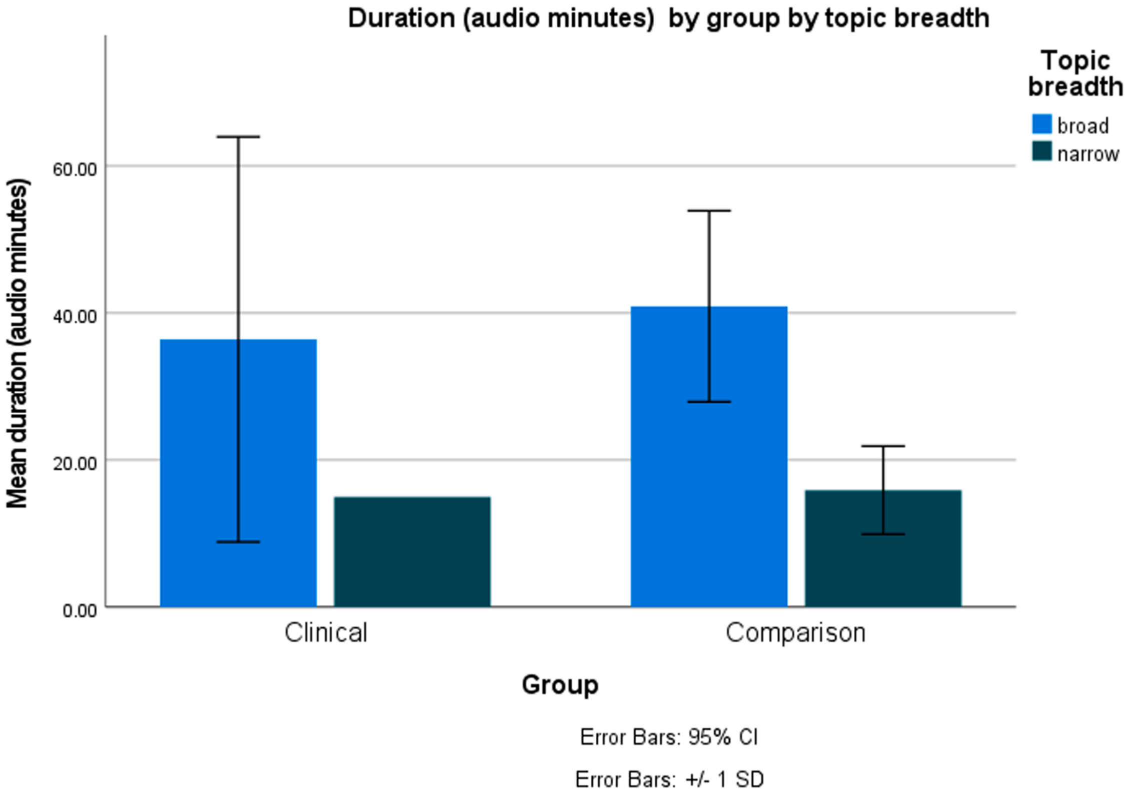


Fig. 12. Duration (audio minutes) by group by topic breadth.

participants also produced a higher token mean on the broad topic than the narrow topic, but the broad topic dispersion is much greater. Comparison participants addressing the narrow topic showed the smallest dispersion overall. Readers are reminded of the following: comparison broad ($n = 4$), comparison narrow ($n = 10$), clinical broad ($n = 14$), clinical narrow ($n = 1$).

Fig. 12 below shows the mean duration in audio minutes by group by topic breadth.

Comparison participants show a higher duration mean on the broad topic than the narrow topic and the highest duration mean overall, despite overrepresentation of the narrow topic in the comparison group and overrepresentation of the broad topic in the clinical group. Comparison dispersions were closer to the comparison means. Clinical participants also produced a higher duration mean on the broad topic, but the broad topic dispersion is much greater. Readers are reminded of the following: comparison broad ($n = 4$), comparison narrow ($n = 10$), clinical broad ($n = 14$), clinical narrow ($n = 1$).

Fig. 13 below shows the mean % of sub-corpus by group by topic breadth.

Clinical and comparison participants show the highest sub-corpus contributions in the broad genre, with clinical participants contributing more to their respective sub-corpus per speaker than the comparison group. Comparison participants show the highest sub-corpus contributions in the broad genre, despite overrepresentation of the narrow topic in the comparison group. The dispersion is much larger for the clinical group. Clinical participants show higher sub-corpus contributions than the comparison group for the narrow genre, as well, with the comparison group producing the smallest dispersion. Readers are reminded of the following: comparison broad ($n = 4$), comparison narrow ($n = 10$), clinical broad ($n = 14$), clinical narrow ($n = 1$).

Fig. 14 below shows the mean % of corpus by group by topic breadth.

Comparison participants show the highest contributions in the broad genre, despite overrepresentation of the narrow genre in this cohort. This is also one of only two cases (the other being mean % of corpus by sex) where dispersion is greater in the comparison group than the clinical group. Clinical group participants show higher contributions in the broad genre than the narrow genre. Comparison group participants show higher contributions in the narrow genre than the clinical participants in the narrow genre. Dispersion was smallest in the comparison narrow genre Readers are reminded of the following: comparison broad ($n = 4$), comparison narrow ($n = 10$), clinical broad ($n = 14$), clinical narrow ($n = 1$).

Broad genre comparison participants produced more tokens, spoke for longer, contributed more to their respective sub-corpus, and contributed more to the corpus overall than narrow genre comparison participants despite being situationally underrepresented at a ratio of 1:3.6. Comparison participants showed the greatest variance for% of corpus by sex and% of corpus by topic breadth.

Summary

The DAIS-C was built with close reference to best practices in the development of spoken and specialised corpora. Design issues were expected on the topics of internal versus external building criteria, sampling and selection bias, and recruitment factors (pandemic aside). Reasoned attempts at mitigation followed, and a review of corpus characteristics suggest that they were generally successful. Demographic overrepresentation issues, although inconvenient, do not compromise the data collected so far. They simply limit the extent of viable analyses and their conclusions. These issues can be resolved in time with corpus expansion work.

A review of group-level data suggests that overrepresentation has had little effect on mean token count, mean duration, mean contributions to respective sub-corpora, or mean contributions to the corpus as a whole. On average, female clinical participants produced more tokens, spoke for longer, contributed more to their respective sub-corpus, and contributed more to the overall corpus than clinical males despite being demographically underrepresented at a ratio of 1:4. Broad genre

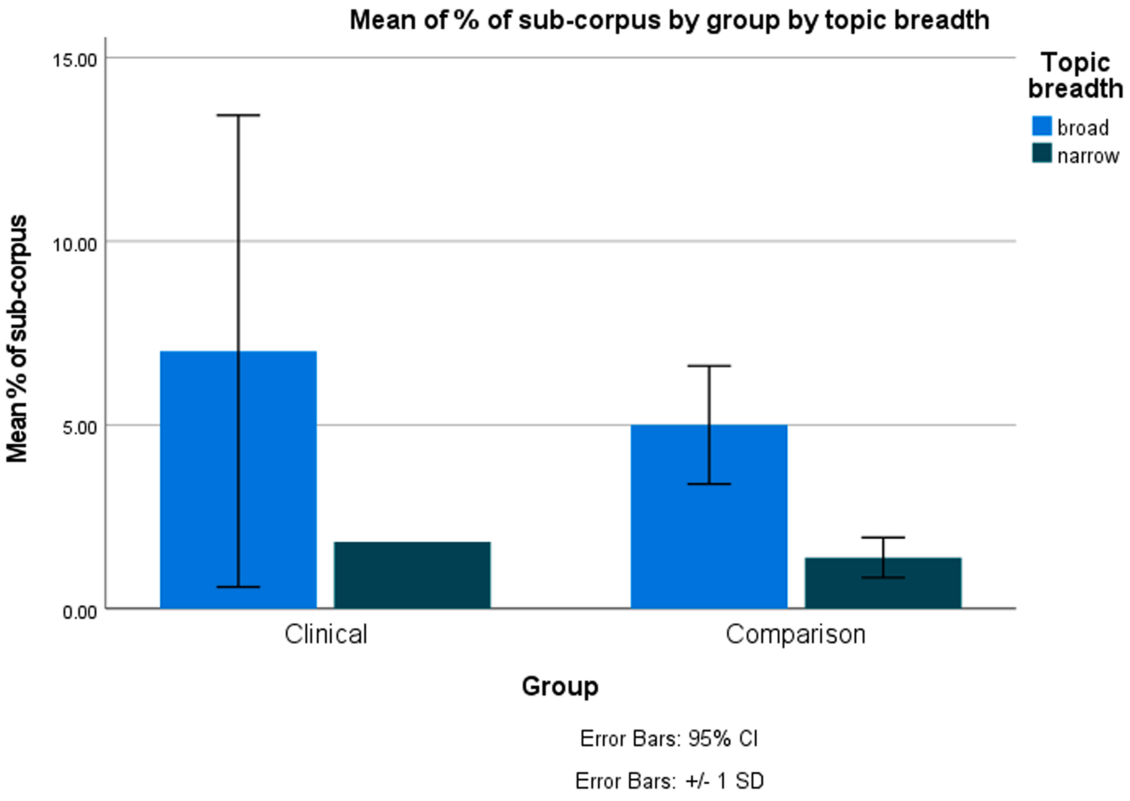


Fig. 13. Mean % of sub-corpus by group by topic breadth.

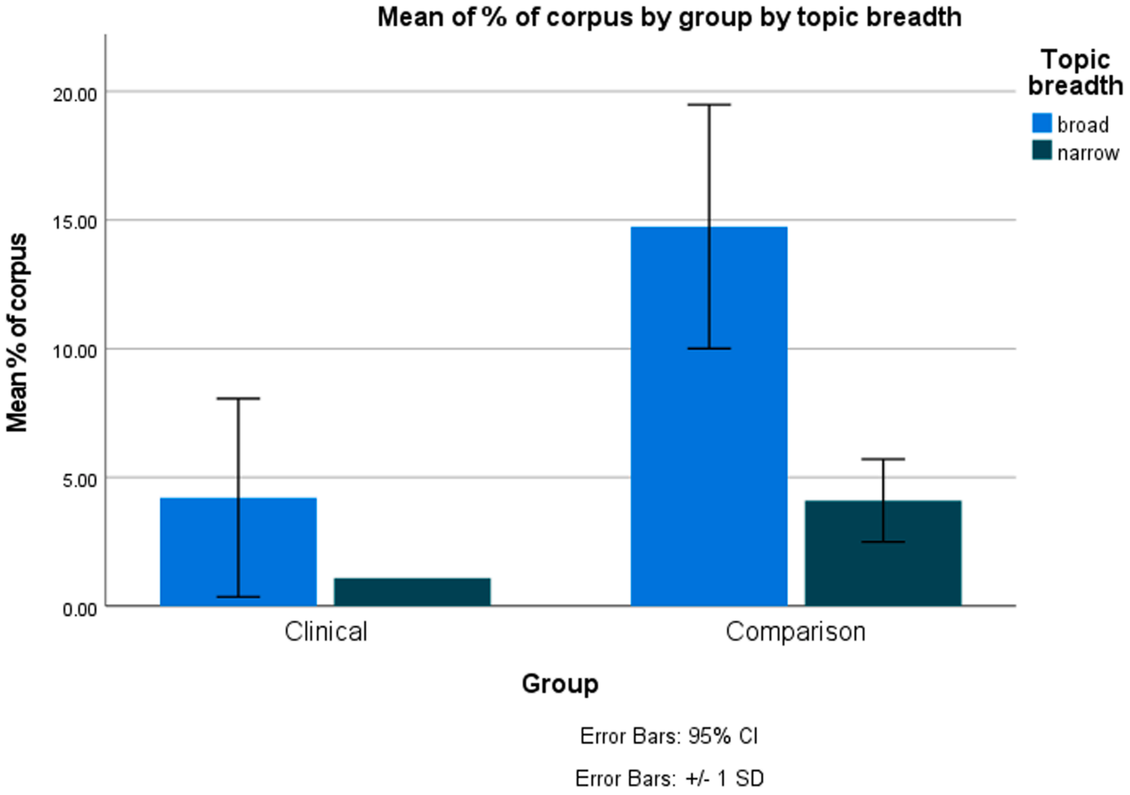


Fig. 14. Mean % of corpus by group by topic breadth.

comparison participants produced more tokens, spoke for longer, contributed more to their respective sub-corpus, and contributed more to the corpus overall than narrow genre comparison participants despite being situationally underrepresented at a ratio of 1:3.6. Female clinical participants also showed the least variance across all sex indices. Comparison participants showed the greatest variance for% of corpus by sex and % of corpus by topic breadth.

Availability

Repository

The UK Data Service has confirmed acceptance of the corpus for archival through its ReShare repository.

Release

Version 1.0 of the DAIS-C will be made available in late 2023.

Applications

The corpus was designed, transcribed, annotated, and stored with multiple future applications in mind. Raw text versions of the corpus are ready for automatic tagging, and pre-annotated versions allow for supplementary analyses beyond those already carried out by the corpus creators. The corpus can be integrated into programmatic or software-based workflows, as speaker labels and interview turns are already bracketed by XML tags. Users wanting to expand this approach can do so by beginning with the XML versions of each speaker file or the full interview file. Sentiment analysis is also possible, as each speaker's responses are stored as individual text files, without interview contributions, in plain text or XML formats, ready for input into software that works file by file. Files can be combined easily either programmatically or with software designed for this purpose, like TXTCollector. It is not possible to list all potential applications of the corpus, but those working in the medical humanities may benefit from comparing clinical and comparison speakers to determine patterns of potential diagnostic or predictive importance. The presence of speakers with and without a history of FTD also allows for some further comparison, although only a handful of speakers in the clinical group meet this criterion. Those working in machine learning might find the corpus helpful as a source of training data, for instance. The corpus is also potentially suited to some qualitative analyses, although it was not designed for this. Work that does not require strict adherence to semi-structured interviewing practices is more likely to benefit from this corpus. While it is not possible to predict how useful this corpus will be to other researchers, this paper nonetheless demonstrates that it is possible to build corpora with reuse in mind. Corpus builders are encouraged to attempt this where feasible given the relative lack of corpora designed according to best practices in corpus linguistics.

Declaration of Competing Interest

We declare no competing interest.

References

Abu-Akel, A., Caplan, R., Guthrie, D., Komo, S., 2000. Childhood schizophrenia: responsiveness to questions during conversation. J. Am. Acad. Child Adolesc. Psychiatry 39 (6), 779–786.

Adolphs, S., Knight, D., 2010. Building a spoken corpus. The Routledge Handbook of Corpus Linguistics. Routledge, pp. 38–52.

Almut, K., 2010. Building small specialised corpora. TheRoutledge Handbook of Corpus Linguistics. Routledge, pp. 66–79.

Andreasen, N.C., 1982. Negative symptoms in schizophrenia: definition and reliability. Arch. Gen. Psychiatry 39 (7), 784–788.

Atkins, S., Clear, J., Ostler, N., 1992. Corpus design criteria. Lit. Linguist. Comput. 7 (1), 1–16.

Batinić, J., Frick, E., Schmidt, T., 2021. Accessing spoken language corpora: an overview of current approaches. Corpora 16 (3), 417–445.

Biber, D., 1990. Methodological issues regarding corpus-based analyses of linguistic variation. Lit. Linguist. Comput. 5 (4), 257–269.

Biber, D., 1993. Representativeness in corpus design. Lit. Linguist. Comput. 8 (4), 243–257.

BNC Consortium, 2007. The British National Corpus, XML Edition. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554.

Burnard, L., 2005. Metadata for corpus work. Developing Linguistic Corpora: A Guide to Good Practice. Oxbow Books, pp. 30–46.

Carter, R., Mncarthy, M., 1995. Grammar and the spoken language. Appl. Linguist. 16 (2), 141–158.

Carter, R., 2015. Language and Creativity: The Art of Common Talk. Routledge.

Clayton, J.A., Tannenbaum, C., 2016. Reporting sex, gender, or both in clinical research? JAMA 316 (18), 1863–1864.

Cook, G., 1990. Trancribing infinity: problems of context presentation. J. Pragmat. 14 (1), 1–24.

Davies, M., 2015, "Corpus of contemporary American English (COCA)", https://doi.org/10.7910/DVN/AMUDUW, Harvard Dataverse, V2.

Delgaram-Nejad, O., Chatzidamianos, G., Archer, D., Larner, S., 2020. What is linguistic creativity in schizophrenia? J. Interact. Res. Commun. Disord. 11 (2), 194–216.

Delgaram-Nejad, O., Chatzidamianos, G., Archer, D., Bartha, A., Robinson, L., 2022. A tutorial on norming linguistic stimuli for clinical populations. Applied Corpus Linguistics 2, 100022.

Du Bois, J.W., 1992. Discourse transcription. Santa Barbara papers in linguistics, 4. University of California, Santa Barbara, California, pp. 1–225.

Edwards, J.A., 1993. Principles and contrasting systems of discourse transcription. Talking Data: Transcription and Coding in Discourse Research. Lawrence Erlbaum Associates, pp. 3–31.

Fervaha, G., Takeuchi, H., Foussias, G., Agid, O., Remington, G., 2016. Using poverty of speech as a case study to explore the overlap between negative symptoms and cognitive dysfunction. Schizophr. Res. 176 (2–3), 411–416.

Flowerdew, L., 2004. The argument for using English specialized corpora to understand academic and professional language. Discourse in the Professions: Perspectives from Corpus Linguistics. John Benjamins, pp. 11–33.

Flowerdew, L., 2014. Corpus-based analyses in EAP. Academic Discourse. Routledge, pp. 105–124.

Gabrić, P., Nagels, A., Kircher, T., Rosenkranz, A., 2021. Within-sample, but not corpus-based word frequency of verbal fluency output is associated with positive symptoms in schizophrenia. doi: 10.31234/osf.io/7tndz.

General Medical Council, 2013. Good Practice in Research and Consent to Research. General Medical Council.

Graddol, D., Maybin, J., Stierer, B., 1994. Researching Language and Literacy in Social Context: A Reader. Multilingual Matters.

Halliday, M.A., 2004. The spoken language corpus: a foundation for grammatical theory. Advances in Corpus Linguistics. Brill, pp. 9–38.

Leech, G., Myers, G., Thomas, J., 2014. Spoken English on Computer: Transcription, Mark-Up and Application. Routledge.

Li, C.R., Chung, Y.C., Park, T.W., Yang, J.C., Kim, K.W., Lee, K.H., Hwang, I.K., 2009. Clozapine-induced tardive dyskinesia in schizophrenic patients taking clozapine as a first-line antipsychotic drug. World J. Biol. Psychiatry 10 (4–3), 919–924.

Marengo, J.T., Harrow, M.M., Lanin-Kettering, I., Wilson, A., 1986. Evaluating bizarre-idiosyncratic thinking: a comprehensive index of positive thought disorder. Schizophr. Bull. 12 (3), 497–511.

McCarthy, M., 1998. Spoken Language and Applied Linguistics. Cambridge University Press.

McKenna, P.J., Oh, T.M., 2005. Schizophrenic Speech: Making Sense of Bathroots and Ponds That Fall in Doorways. Cambridge University Press.

McKenna, P.J., 2007. Schizophrenia and Related Syndromes. Routledge.

Mikesell, L., Bromley, E., 2016. Exploring the heterogeneity of 'schizophrenic speech. The Palgrave Handbook of Adult Mental Health: Discourse and Conversation Studies. Springer, pp. 329–351.

Mitkov, R., 2022. The Oxford Handbook of Computational Linguistics. Oxford University Press.

Mouritsen, S.C., 2019. Contract interpretation with corpus linguistics. Wash. L. Rev. 94, 1337.

O'Keeffe, A., 2007. The pragmatics of corpus linguistics. In: Proceedings of the Fourth Corpus Linguistics Conference.

Ochs, E., 1979. Transcription as theory. Dev. Pragmat. 10 (1), 43–72.

Oomen, P.P., de Boer, J.N., Brederoo, S.G., Voppel, A.E., Brand, B.A., Wijnen, F.N., Sommer, I.E., 2022. Characterizing speech heterogeneity in schizophrenia-spectrum disorders. J. Psychopathol. Clin. Sci. 131 (2), 172.

Pizarro Pedraza, A., 2019. MadSex: collecting a spoken corpus of indirectly elicited sexual concepts. Lang. Resour. Eval. 53 (1), 191–207.

Randi, R., 2010. Building a corpus: what are the key considerations? The Routledge Handbook of Corpus Linguistics. Routledge, pp. 31–37.

Sinclair, J., Wynne, M., 2004. Developing linguistic corpora: A guide to good practice. Ahds Literature, Languages and Linguistics. University of Oxford, UK.

Sinclair, J., 2005. Meaning in the framework of corpus linguistics. Lexicographica 20 (2004), 20–32.

Strassel, S., Cole, A.W., 2006. Corpus development and publication. In: Proceedings of LREC. Genoa, Italy. http://papers.ldc.upenn.edu/LREC2006/CorpusDevelopmentAndPublication.pdf.

Thompson, P.A., 2005. Spoken language corpora.

Warren, M., 2004. //so what have YOU been WORking on REcently: compiling a specialized corpus of spoken business English. Discourse in the Professions. John Benjamins, pp. 115–140.

Waterman, A.H., Blades, M., Spencer, C., 2001. Interviewing children and adults: the effect of question format on the tendency to speculate. Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn. 15 (5), 521–531.

Wray, A., Trott, K., Bloomer, A., 1998. Projects in Linguistics: A Practical Guide to Researching Language. Arnold.