BUILDING A CORPUS OF PHRASEOLOGY EXPRESSING STRESS AND ANXIETY: A CASE FOR SYSTEMATIC COLLECTION AND SEMANTIC CLASSIFICATION

Saidova Zulfizar Khudoyberdievna
Associate professor (PhD) of English linguistics department of Bukhara state university
z.x.saidova@buxdu.uz
Ruzmatova Gulasal Gayratovna
Master student of Bukhara international university

Annotation. *This text argues for a systematic, principled approach to corpus compilation and offers a rationale and framework for identifying and classifying the semantic types of such phraseologisms. The goal is to persuade scholars to adopt rigorous methods that maximize the corpus's usefulness for interdisciplinary research and practical applications.*

Key words: *phraseology; idioms; multiword expressions; stress; anxiety; emotional language; corpus linguistics; specialized corpus; semantic classification; conceptual metaphor; pragmatic annotation; interdisciplinary research; psychological discourse; affective expressions; annotation schema; automatic extraction; human-in-the-loop validation; metadata; figurative language; linguistic markers of distress; NLP; clinical discourse analysis; sociolinguistic variation; ethical data handling.*

Phraseology—fixed or semi-fixed expressions such as idioms, collocations, and set phrases—play a crucial role in conveying emotional states. Expressions like "on edge," "beside oneself," or "at breaking point" encapsulate both cognitive and cultural aspects of stress and anxiety. For researchers in linguistics, psychology, and applied fields (e.g., clinical discourse analysis, corpus linguistics), compiling a dedicated corpus of phraseologisms that express stress and anxiety is not merely an archival task; it is a methodological imperative.

Rationale: Why a Specialized Corpus Is Necessary

1. Research precision and granularity A general corpus can reveal frequency patterns of single lexical items, but phraseologisms that index stress and anxiety often rely on multi-word patterns, figurative meanings, and pragmatic cues. Without a specialized corpus, researchers risk under-detecting or misclassifying expressions whose emotional content emerges from fixed phrasing, metaphorical extension, or idiomatic usage.

2. Interdisciplinary utility A purpose-built corpus can support diverse disciplines: clinical psychologists studying linguistic markers of anxiety; sociolinguists examining how communities verbalize stress; computational linguists

developing emotion recognition models; and translators seeking culturally accurate renderings. Uniform annotation and semantic classification enhance interoperability across these domains.

3. Pedagogical and therapeutic applications Language teachers and therapists can use a well-annotated corpus to teach pragmatic competence or to design interventions that help clients recognize and reframe idiomatic expressions associated with distress.

Principles for Corpus Compilation

1. Clear scope and definitional boundaries Begin by operationally defining what counts as a phraseologism for this project. Include idioms, binomials, verb–object collocations, light-verb constructions, phrasal verbs, and recurrent metaphorical expressions whose established or contextually inferred meanings index stress, anxiety, panic, or related affective states. Exclude purely compositional expressions where emotional meaning arises solely from individual words in context (unless recurrent patterning indicates conventionalized phrasing).

2. Diverse and representative sources Gather data from multiple registers and genres to capture variation: social media (microblogs, tweets), spoken corpora (interviews, talk shows, clinical consultations), fiction and drama, news reporting, self-help literature, forums, and transcripts from counseling interactions (with ethical clearance). Balance formal and informal registers and ensure demographic breadth (age, gender, dialects) to capture sociolinguistic variation.

3. Ethical and legal compliance When using human subjects' data or personal posts, secure informed consent where required and anonymize sensitive information. Follow platform terms of service and institutional review board (IRB) requirements. Prioritize public-domain texts or obtain permissions for proprietary material.

4. Robust metadata Annotate each instance with metadata: source type, date, author/speaker (anonymized), register, geographical/dialectal context, and situational notes (if available). Metadata enables stratified analyses (e.g., frequency of a phraseologism in online vs. spoken data).

5. Multi-level annotation Implement layered annotation to capture form, meaning, and function:

- Form: tokenization and lemmatization; mark fixed multiword expressions (MWEs).
- Semantics: sense labels indicating which emotional state is expressed (e.g., unease, acute panic, chronic worry).
- Pragmatics: illocutionary force (complaint, warning, self-report), intensity, and whether the expression is figurative, hyperbolic, or literal.

- Discourse role: whether the phrase functions as an evaluative stance, an emphatic marker, a hedging device, or a symptom-report.

Methodology for Identification

1. Seed lists and bootstrapping Start with a curated seed list of well-known stress/anxiety phraseologisms drawn from dictionaries of idioms, psychological literature (e.g., common symptom descriptions), and pilot corpus searches. Use these seeds to perform concordance searches and extract collocational patterns. Bootstrapping enables discovery of related expressions via distributional similarity and co-occurrence clusters.

2. Automatic extraction techniques Apply computational methods to scale identification:

- N-gram frequency analysis to surface recurrent multiword sequences.
- Association measures (PMI, t-score) to identify strongly collocating word pairs or multiword sequences.
- Pattern-based extraction using regular expressions for known syntactic frames (e.g., “to be X with nerves,” “to have butterflies in the stomach”).
- Distributional semantic models to find paraphrases and related metaphorical expressions by similarity in embedding spaces.

3. Human-in-the-loop validation Automatic extraction should be followed by manual validation by trained annotators to filter false positives and capture nuanced idiomatic meanings. Inter-annotator agreement measures (Cohen’s kappa, Krippendorff’s alpha) should be computed to ensure reliability. Provide annotators with clear coding guidelines and exemplars.

Semantic Classification Framework

A meaningful classification scheme should be theoretically grounded, empirically motivated, and practically useful. The following multi-dimensional framework is recommended:

1. Primary semantic domain Categorize phraseologisms by the broad affective domain they index:

- Acute fear/panic (e.g., “lose one’s nerve,” “freeze up”).
- Chronic anxiety/worry (e.g., “on pins and needles,” “constantly on edge”).
- Stress/tension (e.g., “under the gun,” “stretched thin”).
- Overwhelm/exhaustion (e.g., “burned out,” “at the end of one’s rope”).
- Embarrassment/anticipatory dread (e.g., “dread to think,” “hang on every word”).

2. Figurative source and conceptual metaphor Record the underlying metaphorical mapping (Lakoff & Johnson style) that structures the phrase:

- Pressure-as-force (e.g., “under pressure”).

- Container-as-mind (e.g., "bubbling with anxiety").
- Heat-as-intensity (e.g., "boiling with worry").
- Breakdown-as-damage (e.g., "fall apart"). This enables cross-linguistic and cognitive-comparative analyses.

3. Pragmatic function Classify whether the phrase is:

- Self-reporting (disclosing one's state).
- Evaluative (judging circumstances or others).
- Soothing or minimizing (e.g., euphemisms).
- Hyperbolic intensifier (e.g., "I'm going to explode").
- Social signaling (eliciting empathy, seeking help).

4. Degree and temporality Annotate intensity (mild, moderate, severe) and temporality (momentary vs. persistent). Where possible, link intensity labels to corpus-derived distributional patterns (e.g., co-occurrence with adverbs like "completely," "a little").

5. Conventionality and morphosyntactic fixity Record degree of conventionalization (fully fixed idiom vs. productive phrase) and syntactic properties (fixed word order, allowable substitutions). This helps distinguish true idioms from flexible metaphorical constructions.

Analytic Benefits and Applications

1. Linguistic theory A corpus with precise semantic classification illuminates how languages encode affective states through conventionalized phrasing, supporting theories of idiom formation, metaphor, and construction grammar.

2. Psycholinguistics and clinical practice Researchers can identify linguistic markers that predict anxiety severity or onset. Therapists might use common phraseologisms to understand clients' narratives and cultural idioms of distress, improving culturally competent care.

3. Computational models Annotated data improve natural language processing (NLP) systems for sentiment and mental-health detection by supplying multiword cues and pragmatic labels beyond single-word sentiment lexicons.

4. Translation and intercultural communication Classified phraseologisms aid translators and language learners in rendering authentic, context-appropriate expressions of stress and anxiety across languages.

Implementation Roadmap

1. Project design and pilot Draft a project charter: research questions, scope, ethical approvals, data sources, annotation schema, and timeline. Conduct a pilot with a small heterogeneous dataset to test extraction algorithms and annotation guidelines.

2. Tooling and infrastructure Use corpus management tools (e.g., corpus query systems, annotation platforms). Ensure data storage complies with privacy regulations. Implement version control for annotations and clear documentation.

3. Annotator recruitment and training Recruit annotators with linguistic or psychological training. Provide calibration sessions, detailed manuals, and exemplars. Monitor inter-annotator agreement and refine guidelines iteratively.

4. Iterative refinement Use error analysis to refine extraction patterns and classification categories. Expand the corpus in stratified stages to include underrepresented registers or communities.

5. Dissemination and reuse Publish the corpus and documentation under clear licensing terms. Provide APIs or concordancers for researchers and practitioners. Offer derived datasets (e.g., frequency lists, metaphor inventories) to lower barriers for applied use.

Anticipated Challenges and Mitigations

1. Polysemy and context-dependence Many phraseologisms have context-dependent readings. Mitigation: rely on contextual windows and human validation; mark ambiguous instances for secondary analysis.

2. Register and dialect variation Expressions vary by region and community. Mitigation: ensure balanced sampling and include dialect labels in metadata.

3. Emotional intensity subjectivity Annotators may disagree on intensity labels. Mitigation: create anchored rating scales with empirical exemplars and compute reliability; consider multi-rater averages.

4. Ethical sensitivity Extracting mental-health-related language raises privacy concerns. Mitigation: anonymize data, prioritize public or consented sources, and implement secure data handling.

Compiling a corpus of phraseologisms that express stress and anxiety is both feasible and valuable. It requires careful definitional work, ethical vigilance, mixed-methods extraction and validation, and a multi-dimensional semantic classification scheme. The benefits span theoretical insight, clinical utility, computational robustness, and pedagogical application. Researchers and institutions should commit resources to build and maintain such corpora, prioritize open documentation, and foster interdisciplinary collaboration. By doing so, the academic community will gain a powerful resource that deepens our understanding of how language shapes and reflects mental states—and equips practitioners with better tools for diagnosis, intervention, and cross-cultural communication.

REFERENCES:

1. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

2. Cruse, A. (2011). Meaning in language: An introduction to semantics and pragmatics (3rd ed.). Oxford University Press.

3. Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. MIT Press.

4. Gries, S. T. (2008). Quantitative corpus linguistics with R: A practical introduction. Routledge.

5. Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.

6. Khudoyberdievna, S. Z. (2022). The main features of translation of phraseology from english into uzbek. Scientific Impulse, 1 (3), 523-526.

7. Khudoyberdievna, S. Z. (2021). English phraseology and its integration with terminology. Academicia: An International Multidisciplinary Research Journal, 11(2), 1618-1622.

8. Сайфуллаева Дилафруз Ахмадовна, Мирджанова Наргиза Норкуловна, Саидова Зулфизар Худойбердиевна РАЗВИТИЕ ПРОФЕССИОНАЛЬНЫХ КОМПЕТЕНЦИЙ И ТВОРЧЕСКИХ СПОСОБНОСТЕЙ СТУДЕНТОВ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ // Вестник науки и образования. 2020. №19-2 (97). URL: https://cyberleninka.ru/article/n/razvitie-professionalnyh-kompetentsiy-i-tvorcheskih-sposobnostey-studentov-vysshih-uchebnyh-zavedeniy (дата обращения: 29.11.2025).

9. Khudoyberdievna, S. Z. (2021). English phraseology and its integration with terminology. Academicia: An International Multidisciplinary Research Journal, 11 (2), 1618-1622.

10. Saidova, Z. (2023). Definition of Idioms in Modern Phraseology. ЦЕНТР НАУЧНЫХ ПУБЛИКАЦИЙ (buxdu. uz), 29, 29.

11. Saidova Zulfizar Khudoyberdievna Model training method: classes in the form of buseness games, lessons such as lesson-court, lesson auction, lesson-press Conference // Достижения науки и образования. 2018. №5 (27). URL: https://cyberleninka.ru/article/n/model-training-method-classes-in-the-form-of-buseness-games-lessons-such-as-lesson-court-lesson-auction-lesson-press-conference (дата обращения: 29.11.2025).