AL-HAYAT Research Journal
ISSN (Online): 3006-4767
ISSN (Print):: 3006-4759
Volume 3 Issue 2 Jan-March 2026)

Language of Stress: *A Corpus-Based Study to Detect Early Signs of Suicide through Lexical Choice*

Afshan Ishfaq

Assistant. Professor Head of Academics at Institute of Law, Lahore, afshan.law@pu.edu.pk

Nida Sultan

Lecturer in English at NAMAL, Mianwali, **nida.sultan@namal.edu.pk**

Abstract

Suicide represents one of the most devastating and preventable causes of premature death globally, yet its early detection remains stubbornly elusive. This study advances the hypothesis that language specifically the spontaneous lexical choices individuals make in everyday written and digital discourse constitutes one of the most sensitive and accessible markers of suicidal ideation. We employ corpus-based methodologies to conduct a systematic, quantitative investigation of how the written language of individuals experiencing suicidal ideation differs from that of a matched non-suicidal population. A purpose-built Suicide Discourse Corpus (SDC) of approximately 452,000 tokens was compiled from four heterogeneous sources: anonymized crisis helpline transcripts, Reddit posts from mental health disclosure communities, published first-person narratives of suicidal crises, and archival farewell notes. A Matched Control Corpus (MCC) of 449,800 tokens from general online discourse was constructed as a baseline. Analytical methods include keyness analysis (log-likelihood, G^2), semantic domain profiling using the UCREL Semantic Analysis System (USAS), collocational analysis (Mutual Information scoring), and frequency analysis of grammatical and functional-word categories. Findings reveal a statistically robust lexical signature in suicidal discourse marked by: (a) dramatically elevated pain, suffering, and death-related vocabulary; (b) absolutist and negation-heavy language reflecting cognitive constriction; (c) a depletion of future-oriented temporal reference and positive evaluative terms; (d) heightened first-person singular pronoun use alongside reduced social solidarity vocabulary; and (e) distinctive collocational frames encoding inescapable, internally directed suffering. These patterns align with major psychological theories of suicide including Shneidman's psychache theory, Joiner's Interpersonal Theory of Suicide (IPTS), and Beck's cognitive model of hopelessness. Implications for the design of NLP-assisted early-warning systems, ethical governance of mental health corpus research, and future multilingual extension of this work are discussed at length.

Keywords: *corpus linguistics, suicide prevention, lexical analysis, language of stress, computational stylistics, mental health discourse, psycholinguistics, natural language processing, Keynes, LIWC*

1: Introduction

1.1 Background and Context

Suicide is a complex, multidetermined act that ends more than 700,000 lives each year worldwide and stands as the fourth leading cause of death among individuals aged 15 to 29 (World Health Organization, 2023). For every completed suicide, the WHO estimates that there are as many as twenty attempts, and for every attempt, a far greater number of individuals experience suicidal ideation without ever seeking professional help. The human cost to the individuals who die, to the families they leave behind, and to the communities that absorb the ripple effects of each loss is incalculable.

Despite decades of psychiatric, psychological, and sociological research, reliable early detection of suicidal risk remains one of the most persistent and consequential unsolved problems in mental health. Standard clinical screening tools such as the Columbia Suicide Severity Rating Scale (C-SSRS), the Beck Scale for Suicidal Ideation (BSS), and the Patient Health Questionnaire (PHQ-9) offer structured frameworks for assessing risk when an individual presents to a clinician. However, these instruments share a critical limitation: they require that the individual at risk makes contact with a healthcare professional. Research consistently shows that a substantial proportion estimated at between 20 and 45 percent of those who die by suicide had no contact with mental health services in the year preceding their death, and many had no contact in the month preceding it (Luoma, Martin, & Pearson, 2002; Pirkis & Burgess, 1998). The clinical encounter upon which these tools depend is, for a significant fraction of at-risk individuals, an encounter that never occurs.

This epidemiological reality has prompted researchers to look beyond the consulting room for signals of suicidal distress. Among the most promising of these alternative signals is language. Human beings are compulsive communicators: we externalize our inner worlds through speech, writing, and digital communication in a near-constant stream of discourse. The words we choose the vocabulary we reach for when we are suffering, the grammatical structures that organize our distress, the metaphors through which we make sense of unbearable experience are shaped by, and in turn shape, the psychological states in which we find ourselves. If suicidal ideation leaves distinctive traces in an individual's language, and if those traces are sufficiently systematic and identifiable, they may constitute the basis of a non-clinical, population-level early-warning system of considerable practical value.

The present study investigates this possibility through the lens of corpus linguistics a discipline that applies rigorous quantitative methods to large, systematically compiled collections of naturally occurring language. Corpus linguistics offers a methodology ideally suited to the investigation of psychologically significant language patterns: it is empirical rather than impressionistic, reproducible rather than anecdotal, and capable of detecting subtle, low-frequency patterns that escape the notice of individual clinicians or readers. Applied to the question of suicidal language, corpus methods can reveal what the language of those in psychological extremis actually looks like in aggregate not what clinicians believe it looks like, and not what retrospective memory or selected quotation suggests, but what emerges from systematic, statistical analysis of real texts.

1.2 The Problem of Linguistic Invisibility

One of the most striking features of suicidal ideation from a communicative standpoint is what might be called its linguistic invisibility. Many individuals who are thinking about suicide do not say so directly. They may speak obliquely expressing a wish to disappear, a sense that others would be better off without them, an exhaustion with the effort of being alive without ever using the words suicide, kill, or die. Clinicians describe this phenomenon as indirect or disguised communication of suicidal intent, and it is well documented in the clinical literature (Brown, Beck, Steer, & Grisham, 2000; Fawcett et al., 1990).

From a linguistic perspective, this indirectness is both a challenge and an opportunity. It is a challenge because it means that simple keyword-based approaches scanning text for the word suicide and flagging it

will inevitably miss a large proportion of linguistically expressed suicidal ideation. But it is an opportunity because the indirectness is itself systematic: suicidal individuals reach for a predictable repertoire of circumlocutions, metaphors, and grammatical constructions that, taken together, constitute a distinctive register of psychological extremity. The phrase tired of existing, for instance, does not name death or suicide, but collocational and contextual analysis can identify it as a recurring formulation in suicidal discourse. The diminution of future-oriented language the statistical shrinkage of will, tomorrow, plan, hope does not flag itself as a suicide warning, but frequency analysis reveals it as one of the most reliable discriminators of suicidal from non-suicidal discourse.

It is precisely this kind of systematic but non-obvious linguistic pattern that corpus linguistics is equipped to identify. By compiling sufficiently large corpora of suicidal and non-suicidal discourse and subjecting them to comparative quantitative analysis, we can map the lexical terrain of psychological extremity with a precision and comprehensiveness unavailable to individual clinicians or researchers working with small samples of text.

1.3 Rationale for a Corpus-Based Approach

The application of corpus-linguistic methods to mental health and suicide research is still relatively recent, but it has already produced a body of findings of considerable theoretical and practical significance. Pioneering work by Pennebaker and colleagues on Linguistic Inquiry and Word Count (LIWC) established that function words especially pronouns carry substantial psychological information (Pennebaker, Mehl, & Niederhoffer, 2003). Stirman and Pennebaker's (2001) analysis of the poetry of suicidal and non-suicidal poets demonstrated corpus-detectable differences in death-related vocabulary and social reference. Al-Mosaiwi and Johnstone's (2018) large-scale study of online mental health communities showed that absolutist thinking a feature of suicidal cognition theorized by Shneidman (1993) and others leaves measurable traces in word frequencies.

More recently, the emergence of large social media datasets and the development of powerful natural language processing (NLP) tools have created new possibilities for this research agenda. The CLPsych shared tasks (Coppersmith et al., 2015; Shing et al., 2018; Zirikly et al., 2019) have spurred the development of computational models for suicide risk prediction from online text. Deep learning approaches, including transformer-based models such as BERT and its clinical variants, have been applied to electronic health records and social media with promising results (Ji et al., 2021; Gaur et al., 2019).

However, despite this progress, important methodological gaps remain. The majority of existing studies have been conducted within a computational NLP paradigm that prioritizes predictive performance over linguistic interpretability: they produce classifiers that achieve impressive accuracy but provide limited insight into why certain texts are classified as high-risk. Corpus linguistics, with its emphasis on frequency, keyness, collocation, and semantic prosody, offers a complementary approach that is interpretively richer. By grounding the analysis in established linguistic concepts and statistical measures, a corpus-based study can not only identify risk indicators but also explain what those indicators reveal about the psychological world of suicidal individuals and why they emerge in the language in the ways that they do.

A further rationale for the present study is the need for multi-source, methodologically integrated research. Many existing studies draw on a single data source (one social media platform, one clinical sample, one text type), limiting the generalizability of findings. The present study compiles a corpus from four distinct sources crisis helpline transcripts, social media, published personal narratives, and farewell notes allowing examination of whether the linguistic signature of suicidal ideation is consistent across registers and contexts, and whether source-specific patterns emerge.

1.4 Significance of the Study

The significance of this research extends across three domains: theoretical, methodological, and applied.

Theoretically, the study contributes to the psychological linguistics of mental distress by providing large-scale, corpus-based evidence for the lexical correlates of major psychological theories of suicide. If

Shneidman's psychache theory predicts that suicidal discourse should be saturated with pain vocabulary, and if Joiner's Interpersonal Theory predicts elevated burdensomeness and isolation language, and if these predictions are borne out in systematic corpus analysis, the findings constitute a significant, independent form of validation for these theories validation that is grounded not in self-report scales or clinical interview, but in naturally occurring language behavior. Methodologically, the study demonstrates the value of integrating multiple corpus-linguistic analytical techniques Keynes analysis, semantic domain profiling, collocational analysis, and grammatical frequency analysis in a single investigation of a psychologically significant language variety. This integrative approach allows for a richer and more comprehensive characterization of suicidal discourse than any single method can provide, and it establishes a model that can be replicated in studies of other mental health conditions and in other languages. In applied terms, the findings directly inform the design of NLP-based suicide risk detection systems. By specifying not just which lexemes are overrepresented in suicidal discourse, but the collocational frames and semantic domains in which they occur, the study provides a linguistically principled basis for feature engineering in machine learning models. The identification of indirect lexical routes to suicidal meaning the circumlocutions and semantic displacements documented in the collocational analysis is particularly important for reducing false-negative rates in automated detection systems. The findings also highlight the critical importance of including negative keywords (linguistic absences) alongside positive keywords in detection algorithms.

1.5 Scope and Delimitations

The scope of the present study is defined along several dimensions. Linguistically, the study focuses on the lexical and semantic levels of language analysis, with supplementary attention to grammatical and functional-word patterns. Phonological and prosodic features are not examined, as the data are drawn entirely from written sources. Pragmatic and discourse-level phenomena (e.g., turn-taking in helpline transcripts, coherence and cohesion in narratives) are acknowledged as important but fall outside the scope of the current analysis, which prioritizes breadth of lexical coverage over depth of discourse analysis. In terms of language, the corpus is drawn exclusively from English-language sources. This is a significant delimitation that reflects both the availability of data and the existing research literature, but it also constitutes a limitation: the lexical and grammatical patterns identified may not generalize to other languages, and indeed there is theoretical reason to expect that culturally specific constructions of self, community, and honor will inflect the language of suicidal distress differently in different linguistic communities. The extension of this research agenda to Arabic, Mandarin, Urdu, and other languages is identified as a priority for future work. The study does not aim to develop or validate a specific automated detection tool; rather, it provides the empirical linguistic foundation on which such tools can be built. Questions of algorithmic design, classifier architecture, and system evaluation are left to future work in the NLP and clinical informatics domains. Finally, the study is cross-sectional: the corpus represents a snapshot of suicidal discourse rather than a longitudinal trajectory through suicidal ideation. Longitudinal corpus methods, which would allow tracking of how an individual's language changes as suicidal ideation intensifies or resolves, are methodologically promising but present formidable data-collection challenges; they are discussed in the recommendations for future research.

1.6 Objectives of the Research

1. To identify lexical items significantly overrepresented and underrepresented in suicidal discourse relative to a matched control corpus, using keyness analysis, and to organize these items into theoretically meaningful categories.
2. To map the semantic domain structure of suicidal discourse using the USAS semantic tagger, identifying which domains of meaning are most prominently expanded and contracted in the language of suicidal individuals.
3. To document grammatical and functional-word patterns including pronoun use, temporal deixis, negation, and absolutist language that distinguish suicidal from non-suicidal discourse, and to quantify their effect sizes.

4. To conduct collocational analysis of the highest-frequency stress-related lexemes in the suicidal corpus, revealing the semantic prosodies and associative meaning frames that characterize their use.

1.7 Research Questions

1. What lexical items are statistically key in the Suicide Discourse Corpus relative to the Matched Control Corpus, and what theoretical constructs do the overrepresented and underrepresented keyword sets instantiate?
2. Which USAS semantic domains are most significantly over- and under-represented in suicidal discourse, and what do the patterns of semantic expansion and contraction reveal about the cognitive-emotional world of suicidal individuals?
3. In what ways do suicidal and non-suicidal texts differ in their patterns of pronoun use, temporal reference (past, present, future), negation, and absolutist language, and how large are these differences in terms of standardized effect sizes?
4. What collocational environments do the twenty highest-keyness lexemes occupy in suicidal discourse, and what do their semantic prosodies and collocational frames reveal about the indirect and direct linguistic expression of suicidal ideation?

1.8 Theoretical Framework

The study is guided by a theoretical framework that integrates insights from three intellectual traditions: psychological theories of suicidal behavior, corpus linguistics and computational stylistics, and the psycholinguistics of emotion and cognition. From psychology, the study draws primarily on Shneidman's (1993) psychache theory, which posits unbearable psychological pain as the necessary and sufficient cause of suicide; Joiner's (2005) Interpersonal Theory of Suicide (IPTS), which identifies thwarted belongingness and perceived burdensomeness as the proximal cognitive-affective conditions of suicidal desire; and Beck's (1963, 1967) cognitive model, which emphasizes the negative cognitive triad (negative views of self, world, and future) and hopelessness as central to depression and suicide risk. O'Connor's (2011) Integrated Motivational-Volitional (IMV) model is also incorporated as an integrative framework that maps the motivational and volitional stages of suicidal ideation and provides predictions about how language might shift across these stages.

From corpus linguistics, the study employs the keyness framework (Scott, 1997; Rayson, 2008), which identifies words that are statistically unusual in a target corpus relative to a reference corpus, as the primary discovery procedure. Collocational theory (Firth, 1957; Sinclair, 1991; Hunston, 2002) provides the analytical framework for examining how words combine in revealing ways, and the concept of semantic prosody the tendency of words to occur in contexts that carry a consistent evaluative or attitudinal loading is employed to interpret collocational patterns. Semantic domain analysis using the USAS tagset (Rayson et al., 2004) provides a systematic method for classifying and comparing the meanings deployed in each corpus.

From psycholinguistics, the study draws on research on the linguistic expression of emotion (Kövecses, 2000; Fussell, 2002), the relationship between self-focused attention and first-person pronoun use (Ingram, 1990; Mor & Winquist, 2002), and the role of temporal language in mental representation of past, present, and future (Boroditsky, 2011). This tripartite theoretical framework is not merely a bibliographic gesture: it generates specific, testable predictions about the lexical patterns expected in suicidal discourse, predictions that the corpus analysis is designed to evaluate.

1.9 Organization of the Thesis

The remainder of the paper is structured as follows. Chapter 2 provides a comprehensive review of the theoretical and empirical literature on suicidal language, psychological theories of suicide, and prior corpus and computational studies of mental health discourse. Chapter 3 describes the design of the Suicide Discourse Corpus and the Matched Control Corpus in detail, and explains the analytical methods employed. Chapter 4 presents the findings organized by research question: keyness, semantic domains, grammatical patterns, and collocational profiles. Chapter 5 discusses the findings in relation to the theoretical framework, considers their

implications for NLP-based detection systems, addresses ethical concerns, and acknowledges the limitations of the study. Chapter 6 concludes with a summary of contributions and directions for future research. Full reference lists, corpus composition details, and supplementary frequency tables are provided in the appendices.

It should be noted at the outset that the subject matter of this study the language of individuals in profound psychological distress demands not only methodological rigor but also ethical sensitivity. Throughout the research process, and throughout this report, the authors have endeavored to maintain an approach that treats the individuals whose language forms the corpus not as data points but as human beings whose suffering and, in some cases, whose deaths, deserve to be engaged with respectfully and with full awareness of the research's responsibilities.

2: Review of Literature

2.1 Psychological Theories of Suicide and Their Linguistic Implications

2.1.1 Shneidman's Psychache Theory

Edwin Shneidman's theory of psychache (1993, 1996) remains among the most psycholinguistically generative accounts of suicidal motivation. Shneidman argued that every suicide is driven by unbearable psychological pain, a pain he distinguished carefully from physical pain, social pain, or existential despair, though any of these may contribute to it. Psychache, in his formulation, is the felt inadequacy of psychological needs: the excruciating sense that one's most fundamental needs for love, achievement, control, or understanding cannot be met and will not be met. When this pain crosses the individual's unique threshold of tolerance, suicide appears as the only available exit.

The psycholinguistic implications are direct. If suicidal language is organized around the experience and communication of unbearable psychological pain, we should find elevated frequencies of pain vocabulary in the discourse of suicidal individuals not only the lexeme pain itself, but the entire semantic field of suffering, anguish, agony, torment, and their collocates. Equally predictive is Shneidman's concept of cognitive constriction: his observation that suicidal individuals experience a tunnel-like narrowing of perceived options, coming to believe that there are only two possibilities unending pain or death. Constriction should manifest lexically in the overuse of absolutist terms (nothing, never, always, completely), in reduced lexical diversity, and in the near-absence of conditional or open-ended language.

2.1.2 Joiner's Interpersonal Theory of Suicide (IPTS)

Thomas Joiner's Interpersonal Theory of Suicide (2005) proposes that suicidal desire arises from the co-occurrence of two cognitive-affective states: thwarted belongingness (the perception that one is socially isolated, disconnected, and unloved) and perceived burdensomeness (the belief that one is a burden to others, that one's death would benefit that one love). These states are theorized to give rise to suicidal desire when they are accompanied by the acquired capability for suicide, a reduced fear of death and increased tolerance for physical pain developed through prior exposure to painful or dangerous experiences.

Linguistically, thwarted belongingness predicts elevated frequencies of social isolation vocabulary (alone, lonely, isolated, abandoned, invisible) and a reduction in social solidarity language (we, us, together, belong). Perceived burdensomeness predicts the appearance of self-referential negative evaluations framed in relational terms: the idea that one's existence is harmful to others, expressed through words such as burden, problem, fault, and ruin, frequently in collocations with first-person pronouns (I am a burden, my existence is a problem). Research by Van Orden et al. (2010) and Chu et al. (2017) has provided empirical support for these predictions using LIWC-based text analysis, though corpus-level validation with larger datasets remains sparse.

2.1.3 Beck's Cognitive Model and Hopelessness

Aaron Beck's cognitive model of depression (1963, 1967, 1979) and his subsequent theory of hopelessness as a specific predictor of suicidal behavior (Beck, Weissman, Lester, & Trexler, 1974; Beck, Brown, & Steer, 1989) have generated perhaps the most extensively operationalized predictions for lexical research. The negative cognitive triad characterized by negative views of the self, the world, and the future maps directly onto identifiable lexical fields. Negative self-evaluation words (worthless, useless, failure, pathetic, disgusting) represent the self-directed component; negative world-appraisal words (unfair, cruel, meaningless, pointless, broken) represent the world-directed component; and the collapse of future-oriented language represents the most linguistically distinctive marker of hopelessness.

Beck's Hopelessness Scale (BHS), which explicitly measures negative future expectation, has a transparent lexical analogue: an individual whose discourse is depleted of future temporal reference, conditional constructions, and terms denoting planning, possibility, and anticipation is producing language that encodes the experiential and cognitive state the BHS is designed to measure. This insight suggests that automated analysis of temporal deixis in spontaneous language may serve as a non-reactive, continuous proxy measure for hopelessness an insight with obvious clinical implications.

2.1.4 The Integrated Motivational-Volitional (IMV) Model

Rory O'Connor's Integrated Motivational-Volitional model (2011, 2021) offers a more comprehensive process account of suicidal behavior than any single-factor theory. It distinguishes three phases: a pre-motivational phase shaped by background factors (dispositional, biological, social, and environmental); a motivational phase in which defeat and entrapment the sense of being trapped by circumstances with no escape emerge as the core psychological states; and a volitional phase in which motivational states are converted into suicidal planning and behavior by volitional moderators including impulsivity, social imitation, and access to means.

The IMV model generates phase-specific lexical predictions. In the motivational phase, the dominant lexical markers should be defeat (failed, loser, beaten, crushed) and entrapment (trapped, stuck, no way out, escape); in the volitional phase, language should shift towards more concrete and decided expression of suicidal intent. These predictions suggest that corpus studies capable of tracking an individual's discourse over time as in longitudinal forum data might detect the transition between motivational phases, providing a dynamic rather than static picture of suicide risk.

2.2 Corpus and Computational Studies of Suicidal Language

2.2.1 Foundational LIWC-Based Research

James Pennebaker and colleagues' development of the Linguistic Inquiry and Word Count (LIWC) software programe and its application to psychological research created the methodological foundation for much subsequent work on suicidal language. In their seminal overview, Pennebaker, Mehl, and Niederhoffer (2003) argued that function words—including pronouns, articles, prepositions, and negation markers carry substantial information about the psychological state of their user, and that this information is largely invisible to conscious introspection and strategic management. The finding that depressed and suicidal individuals use I, me, and my at significantly elevated rates was interpreted as reflecting excessive self-focused attention and ruminative self-processing.

Stirman and Pennebaker's (2001) analysis of the poetry of suicidal and non-suicidal poets—using LIWC on a corpus of published work from poets who either died by suicide or did not—remains a landmark study in the field. Suicidal poets used significantly fewer words referring to other people (we, us, friends, society) and significantly more words in the death category. The study was methodologically innovative but limited by small sample size (nine suicidal and nine non-suicidal poets) and the specialised literary genre of the data.

2.2.2 Social Media and Large-Scale Computational Studies

The emergence of social media platforms—and the vast volumes of self-expressive writing they generate—has transformed the landscape of suicidal language research. Coppersmith, Dredze, and Harman (2014) demonstrated that Twitter language from users who had self-disclosed a diagnosis of depression showed characteristic linguistic differences from matched controls, including elevated negativity, altered temporal patterns, and reduced social engagement. Coppersmith et al. (2015) extended this methodology to a range of mental health conditions and to Twitter language from users who had self-disclosed a suicide attempt.

The CLPsych shared tasks, organized around a Reddit-derived dataset of posts from users with documented histories of suicidal crisis, have generated a substantial body of computational research. Shing et al. (2018) reported an annotated corpus of Reddit posts stratified by suicide risk level, which has become a standard benchmark. Ji et al. (2021) applied hierarchical attention networks to this dataset, demonstrating state-of-the-art classification performance but providing limited linguistic interpretability of the learned features.

Al-Mosaiwi and Johnstone (2018), whose work is directly relevant to the present study, used large forum datasets to demonstrate that absolutist thinking operationalized through word lists was a stronger discriminator of suicidal ideation from depression and anxiety than negative emotion alone. This finding has significant theoretical implications: it suggests that the cognitive-structural feature of suicidal thinking (its all-or-nothing, non-contingent character) leaves a more distinctive lexical trace than its emotional valence (its negativity), and that detection systems relying on sentiment alone will miss an important signal.

2.2.3 Clinical and Narrative Text Studies

Beyond social media, researchers have examined suicidal language in clinical and narrative contexts. Gkotsis et al. (2017) applied NLP methods to anonymized electronic health records, finding that clinical notes pertaining to patients with documented suicidal ideation contained distinctive lexical markers including hopelessness expressions, social isolation vocabulary, and references to self-harm history. The study demonstrated that NLP can extract clinically meaningful information from unstructured clinical text, though the specialized vocabulary and genre conventions of clinical writing differ substantially from spontaneous personal discourse.

Research on farewell notes documents that provide unusually direct access to the mental state of suicidal individuals at the moment of maximum crisis has a long history in suicidology. Leenaars (1988) and Schnyder et al. (1999) conducted qualitative and quasi-quantitative analyses identifying themes of hopelessness, burdensomeness, and cognitive constriction. More recent corpus-based analyses (Pestian et al., 2012) have sought to identify the features that distinguish genuine farewell notes from simulated ones, with findings consistent with the broader literature on suicidal language.

2.3 Language, Metaphor, and the Expression of Suicidal Experience

An important strand of research, grounded in cognitive linguistics rather than corpus stylistics, has examined the metaphorical structures through which suicidal individuals conceptualize their experience. Tay (2012) and Priel, Mitrany, and Shahar (1998) document the prevalence of entrapment metaphors (being in a cage, having no door, being buried alive) and pain metaphors (carrying an unbearable weight, being crushed, burning from the inside) in the discourse of suicidal individuals. These metaphors are not merely decorative; they organize the experiential and cognitive structure of suicidal distress in ways that have direct implications for intervention.

For corpus-based research, these findings suggest that attention to collocational patterns around pain and entrapment vocabulary may reveal the metaphorical frames within which suicidal distress is encoded and that metaphorical expressions of suicidal intent (I feel trapped with no way out; this weight is too much to carry) may be as linguistically significant as literal ones (I want to die).

2.4 Gaps in the Existing Literature

Despite the rich and growing literature reviewed above, several important gaps motivate the present study. First, most existing corpus-based studies have relied on a single data source, limiting the generalizability of findings. Second, there is relatively little integration of multiple analytical methods within a single study: LIWC-based research captures broad lexical categories but not collocational structure; computational NLP studies achieve high predictive accuracy but limited linguistic interpretability; qualitative studies provide rich interpretation but lack statistical rigor. Third, the study of lexical absences words that are conspicuously missing from suicidal discourse has received less attention than the study of overrepresented vocabulary, despite the theoretical importance of such absences (depleted future language, reduced social solidarity). Fourth, the ethical dimensions of corpus construction and use in this domain have not been systematically addressed in the published literature. The present study addresses each of these gaps.

3: Corpus Design and Methodology

3.1 Corpus Compilation

A purpose-built Suicide Discourse Corpus (SDC) was compiled from four sources spanning multiple registers and contexts of suicidal writing. This multi-source design was adopted to ensure that findings are not artefacts of any single genre, platform, or population.

First, anonymized transcripts of online crisis helpline exchanges were obtained under a formal data-sharing agreement with a mental health charity, subject to full ethics approval and anonymization procedures. Second, posts from the Reddit communities r/Suicide Watch and r/depression were collected using the Pushshift API for the period January 2018 to December 2022, restricted to posts in which users explicitly disclosed suicidal ideation or a current attempt. Third, published first-person narratives of suicidal crises drawn from clinical memoirs and academic anthologies were digitized and incorporated. Fourth, a collection of farewell notes compiled from publicly available suicidology research archives was added.

A Matched Control Corpus (MCC) of approximately equivalent size was constructed from posts in the general-audience Reddit communities r/Ask Reddit and r/off my chest, screened rigorously to exclude any mention of suicide, self-harm, or severe mental illness. Control texts were matched to SDC texts on approximate text length, platform, and time period.

Source	Tokens (SDC)	Documents	Notes
Crisis helpline transcripts	87,450	312	Fully anonymized; ethics-approved
Reddit – r/Suicide Watch	163,200	891	Posts disclosing active ideation
Reddit – r/depression	121,300	702	Posts mentioning suicidal thoughts
Personal narratives / memoirs	52,600	44	Published; public domain
Farewell notes	27,450	118	Archival suicidology data
Total SDC	452,000	2,067	
Total MCC (control)	449,800	2,047	Matched for length and platform

Table 1. Composition of the Suicide Discourse Corpus (SDC) and Matched Control Corpus (MCC)

3.2 Corpus Processing

All corpus texts were processed through a standardized NLP pipeline. Raw HTML and metadata were stripped; texts were sentence-tokenized, word-tokenized, and lemmatized using spaCy (version 3.5) with the en_core_web_lg model. Part-of-speech tags were assigned using spaCy's statistical tagger. Semantic tags were applied using the UCREL Semantic Analysis System (USAS) Python API. Proper nouns and usernames were replaced with generic placeholder tokens to protect anonymity. Texts were retained only if they were at least 50 tokens long to avoid unreliable frequency estimates from very short documents.

3.3 Analytical Methods

3.3.1 Keyness Analysis

Keyness analysis was used as the primary discovery procedure for identifying lexical items significantly more or less frequent in the SDC relative to the MCC. The log-likelihood statistic (G^2) was employed as the keyness measure, with a significance threshold of $p < 0.001$ ($G^2 \geq 10.83$). Effect size was measured using the %DIFF metric (Gabrielatos & Marchi, 2012). Both positive keywords (significantly overrepresented in the SDC) and negative keywords (significantly underrepresented) were retained and analyzed.

3.3.2 Semantic Domain Analysis

USAS semantic tags were aggregated at major and minor category levels and compared between corpora using chi-square tests with Bonferroni correction for multiple comparisons. Proportions of tagged tokens in each semantic domain were calculated and contrasted, with significant differences reported as relative percentage differences.

3.3.3 Collocational Analysis

Collocations of the twenty highest-keyness lexemes were extracted using a ±4-word span and ranked by Mutual Information (MI) score, with a minimum frequency threshold of 5 co-occurrences. MI scores above 3.0 were considered meaningful. The semantic prosodies of key collocational profiles were interpreted qualitatively by two independent researchers, with disagreements resolved by discussion.

3.3.4 Grammatical and Functional-Word Analysis

Normalized frequencies per 1,000 tokens were calculated for first-person singular pronouns, first-person plural pronouns, negation tokens, absolutist words, past-tense verbal forms, future-tense markers, and death- and pain-related lemma clusters. Group comparisons were made using independent-samples t-tests with Welch's correction for unequal variance; effect sizes were reported as Cohen's d.

3.4 Ethical Framework

The study received ethics approval from the Institutional Review Board (Protocol #XXXX-XXXX). Online data collection complied with Reddit's API terms of service and with GDPR data minimisation principles. All usernames were pseudonymized or redacted. Crisis helpline data were shared under strict contractual data-use conditions. The corpus is not publicly distributed. The study adopts a principle of ethical minimalism: no more data was collected or retained than was necessary to answer the research questions. The research team acknowledges the tension between the public health benefits of this research and the privacy interests of the individuals whose language forms the data, and discusses this tension explicitly in Chapter 5.

4: Findings

4.1 Keyness Analysis: Positive and Negative Keywords

Keyness analysis yielded 1,847 positive keywords and 1,213 negative keywords at $p < 0.001$. Table 2 presents the top twenty positive keywords with frequencies and G^2 values.

Rank	Keyword (lemma)	SDC /1k	MCC /1k	G^2	Semantic Category
1	want	28.4	14.1	1,842.3	Desire / volition
2	feel	26.7	11.2	1,791.5	Affect / cognition
3	pain	19.3	4.2	1,654.2	Psychological pain
4	nothing	17.8	6.9	1,489.7	Absolutism / negation
5	never	16.2	7.4	1,301.8	Absolutism / negation
6	alone	15.9	4.8	1,288.4	Social isolation
7	hurt	14.7	4.1	1,201.3	Pain / suffering
8	hate	13.8	5.6	1,154.6	Negative emotion
9	tired	13.4	4.7	1,089.2	Exhaustion / depletion
10	anymore	13.1	3.2	1,072.1	Temporal finality
11	die	12.9	1.8	1,061.3	Death / finality
12	burden	11.8	1.1	998.4	Perceived burdensomeness
13	worthless	11.6	1.4	986.7	Negative self-evaluation
14	no one	11.2	3.8	934.5	Social isolation / absolutism
15	end	10.9	5.1	901.2	Finality / termination
16	hopeless	10.7	0.9	898.6	Hopelessness
17	kill	10.3	2.7	867.1	Death (self-directed)
18	empty	9.8	2.1	831.9	Psychological numbness
19	always	9.6	5.9	801.4	Absolutism
20	escape	9.1	2.3	789.3	Entrapment / relief-seeking

Table 2. Top 20 positive keywords in the SDC (all significant at $p < 0.001$, G^2 threshold 10.83)

The pattern is theoretically coherent and striking in its consistency. Pain, suffering, death, absolutism, social isolation, and negative self-evaluation cluster at the top of the keyword list. Among the top twenty keywords, five belong to the absolutism/negation cluster, four index pain and physical-psychological depletion, three directly reference death and finality, and two (burden, worthless) maps directly onto Joiner's constructs of perceived burdensomeness and negative self-evaluation. The presence of escape (rank 20) is theoretically significant, reflecting Baumeister's (1990) escape theory and the IMV model's entrapment construct.

The top negative keywords are equally revealing: laugh, funny, weekend, game, excited, plan, future, project, learn, tomorrow, and friend (as positive social agent). Their disappearance from the SDC paints a

complementary portrait: suicidal discourse is a world from which humor, curiosity, anticipation, social pleasure, and temporal openness have withdrawn.

4.2 Semantic Domain Analysis

Dir.	Semantic Domain (USAS)	SDC %	MCC %	χ^2 (sig.)
▲	E4.1+ Negative emotion / sadness	8.42	3.18	$p < 0.0001$
▲	E6 Worry, anxiety, fear	6.31	2.41	$p < 0.0001$
▲	B1 Life / death (mortality)	5.87	1.34	$p < 0.0001$
▲	X9.1 Thought / certainty / absolutism	5.14	2.89	$p < 0.0001$
▲	S1.2 Relationship: social exclusion	4.62	1.52	$p < 0.0001$
▼	S1.1 Relationship: friendship/solidarity	1.21	4.78	$p < 0.0001$
▼	K5 Entertainment / leisure	0.84	3.91	$p < 0.0001$
▼	A9+ Positive evaluation / assessment	1.14	5.23	$p < 0.0001$
▼	T1.3 Time: Future reference	1.89	4.47	$p < 0.0001$
▼	Q2.2 Learning / knowledge / curiosity	0.91	3.12	$p < 0.0001$

Table 3. USAS semantic domain comparison. ▲ = overrepresented in SDC; ▼ = underrepresented.

The depletion of future temporal reference (T1.3: 1.89% in SDC vs. 4.47% in MCC) is among the most theoretically significant findings. This temporal collapse, the statistical shrinkage of the future directly instantiates Beck's hopelessness construct and Shneidman's cognitive constriction at the corpus level.

4.3 Grammatical and Functional-Word Patterns

Feature	SDC /1k	MCC /1k	t-value	Cohen's d
1st-person singular (I, me, my, myself)	89.4	61.2	t = 28.4***	0.74
1st-person plural (we, us, our)	12.1	24.6	t = −19.8***	−0.52
Negation tokens	38.9	21.4	t = 31.7***	0.83
Absolutist words	22.7	12.3	t = 22.1***	0.58
Past tense ratio	0.64	0.51	t = 18.3***	0.48
Future tense forms	8.2	16.7	t = −24.6***	−0.64
Death-related lemmas	14.3	2.8	t = 41.2***	1.08
Pain/suffering lemmas	18.6	3.4	t = 48.7***	1.27

*Table 4. Grammatical and lexical category comparisons (***$p < 0.001$). Effect sizes: $d > 0.8$ = large.*

Pain/suffering lemmas (d = 1.27) and death-related lemmas (d = 1.08) show the largest effect sizes in the study—large effects by conventional standards (Cohen, 1988). Negation (d = 0.83), first-person singular

pronouns (d = 0.74), and depletion of future forms (d = −0.64) are also substantial. The temporal pattern is unambiguous: suicidal discourse is past-saturated and future-depleted—anchored in a painful past and present from which the future has largely disappeared.

4.4 Collocational Profiles of Key Lexemes

4.4.1 PAIN

In the SDC, pain collocates most strongly with unbearable (MI = 9.4), emotional (MI = 8.7), constant (MI = 8.2), stop (MI = 7.9), and numb (MI = 7.6). The dominant collocational frame constructs pain as a continuous, overwhelming inner experience from which relief is desired. In the MCC, pain collocates with physical, muscle, knee, and chronic—indexing bodily rather than psychological suffering. The different semantic prosodies are striking: pain in the SDC occupies a frame of psychological, permanent, intolerable suffering; in the MCC, it is contingent, bodily, and treatable.

4.4.2 ALONE

In the SDC, alone collocates with completely (MI = 9.1), feel (MI = 8.8), all (MI = 8.3), left (MI = 8.0), and truly (MI = 7.7). The phrase feel completely alone is among the most characteristic collocations in the entire corpus. The absolutist modifier completely marks the cognitive constriction dimension of perceived isolation. In the MCC, alone collocates with home, sit, time, and enjoy—voluntary solitude rather than involuntary abandonment.

4.4.3 TIRED

Tired in the SDC collocates with so (MI = 8.9), fighting (MI = 8.5), everything (MI = 7.8), pretending (MI = 7.4), and—most significantly—existing (MI = 7.2). The collocation tired of existing directly expresses the desire to cease to be without naming suicide: a form of semantic displacement that circumvents direct reference to death while encoding it. This finding has important implications for detection systems: surface keyword spotting that only scans for die or suicide will miss this construction.

4.4.4 END

End in the SDC collocates with it (MI = 8.3), want (MI = 8.1), everything (MI = 7.9), life (MI = 7.7), and just (MI = 6.4). The cluster want to end it / want everything to end constructs a desire for total termination. In the MCC, end collocates with weekend, story, game, year—entirely different semantic fields. The same word carries fundamentally different meaning worlds in the two corpora.

5: Discussion

5.1 Theoretical Implications

The findings provide corpus-level empirical support for multiple leading psychological theories of suicide. Pain, suffering, and relief vocabulary confirm psychache theory's prediction that suicidal discourse is organized around unbearable psychological pain. Cognitive constriction is confirmed by the absolutism and negation findings. Perceived burdensomeness and thwarted belongingness vocabulary confirm Joiner's IPTS predictions. The temporal collapse past-saturated, future-depleted discourse directly instantiates Beck's hopelessness construct. The presence of escape and entrapment vocabulary confirms the IMV model's prediction of these features in the motivational phase. Taken together, the corpus findings constitute convergent, independent, linguistic-level validation for each of these theories.

Perhaps most theoretically significant is the finding regarding tired of existing. This collocation represents a class of indirect lexical expressions of suicidal intent semantic displacements that encode the desire for death without directly naming it. These displacements are linguistically systematic (they recur at predictable rates, with predictable collocates) and theoretically meaningful (they reflect the taboo character of explicit suicide speech). Their identification has both theoretical value (demonstrating the linguistic creativity

with which suicidal distress is expressed) and practical importance (requiring detection systems to move beyond keyword matching).

5.2 Implications for NLP-Based Early Detection

The findings suggest several concrete design principles for NLP-assisted suicide risk detection systems. First, the large effect sizes for pain and death vocabulary indicate that these domains should receive high weighting in any feature set. Second, the importance of absolutist language identified by Al-Mosaiwi and Johnstone (2018) and confirmed here suggests that sentiment alone is insufficient and that cognitive-structural features must be incorporated. Third, negative keywords (temporal depletion, social solidarity absence, positive evaluation absence) provide a complementary signal that may catch cases where explicit distress vocabulary is absent but linguistic withdrawal is evident. Fourth, collocational context is essential: the same word (end, alone, tired) carries entirely different meanings in suicidal and control discourse, and systems that ignore context will generate high false-positive rates.

5.3 Ethical Considerations

The deployment of language-based suicide detection raises profound ethical questions. Four are identified as primary. First, privacy and consent: individuals writing in online forums have not consented to linguistic surveillance. Research and detection uses of such data require institutional ethics oversight, robust anonymization, and ongoing engagement with questions of proportionality and purpose limitation. Second, algorithmic harm: false positives may trigger stigmatizing or coercive interventions; false negatives may provide false reassurance. High-stakes deployment requires rigorous evaluation and transparent governance. Third, epistemic justice: the present corpus is English-language and platform-specific; deployment of models trained on these data in multilingual or non-Western contexts would be unjustified without replication. Fourth, surveillance scope: the difference between a trained clinician noticing linguistic distress in a conversation and an automated system scanning millions of communications continuously is qualitative, not merely quantitative. Any such system must be embedded in appropriate regulatory, ethical, and human-oversight frameworks.

5.4 Limitations

Four principal limitations should be acknowledged. First, the corpus consists exclusively of individuals who disclosed their suicidal ideation; those who do not disclose may show different linguistic patterns. Second, the data are English-language and Reddit-skewed, limiting generalizability. Third, the USAS tagger's accuracy may be reduced on informal, emotionally charged online language. Fourth, the study is cross-sectional; longitudinal data tracking linguistic change across the arc of suicidal crisis would be considerably more powerful.

6: Conclusion

This study has demonstrated, through rigorous corpus-based analysis of a purpose-built multi-source corpus of over 900,000 tokens, that suicidal discourse carries a distinctive and statistically robust lexical signature. This signature encompasses elevated frequencies of pain and suffering vocabulary, absolutist and negation-heavy language, death and finality terms, and social isolation expressions, alongside the conspicuous depletion of future-oriented, positive-evaluative, and social solidarity language.

The collocational analysis reveals that these words are embedded in semantic frames of inescapable, total, internally directed suffering frames whose meaning extends far beyond what surface-level keyword spotting can detect. Indirect lexical routes to suicidal meaning, of which tired of existing is a paradigm case, constitute a dimension of suicidal discourse that has been insufficiently studied and that has major implications for detection system design.

The findings contribute to theory by providing corpus-level evidence for psychache, cognitive constriction, thwarted belongingness, perceived burdensomeness, and hopelessness. They contribute to methodology by demonstrating the value of integrating keyness, semantic domain analysis, collocational analysis, and grammatical frequency analysis in a single investigation. They contribute to practice by specifying a linguistically principled feature set for NLP-based risk detection systems. And they contribute to ethics by articulating the tensions and responsibilities that attend corpus-based research on the language of psychological extremity.

Future research should extend this work to non-English languages; develop longitudinal corpus methods that can track linguistic change across the trajectory of suicidal ideation; conduct participatory research with people with lived experience of suicidal crisis; and develop and validate multi-feature classifiers incorporating the patterns identified here. Language, as this study has sought to demonstrate, is not merely a symptom of psychological distress—it is the medium through which distress is constituted, communicated, and, potentially, heard. The rigorous, ethically grounded study of language may prove to be one of our most powerful tools for reaching those who are suffering in silence.

References

- Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529–542.
- Baumeister, R. F. (1990). Suicide as escape from self. Psychological Review, 97(1), 90–113.
- Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. Archives of General Psychiatry, 9(4), 324–333.
- Beck, A. T. (1967). Depression: Clinical, experimental, and theoretical aspects. Harper & Row.
- Beck, A. T., Brown, G., & Steer, R. A. (1989). Prediction of eventual suicide in psychiatric inpatients by clinical ratings of hopelessness. Journal of Consulting and Clinical Psychology, 57(2), 309–310.
- Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. Journal of Consulting and Clinical Psychology, 42(6), 861–865.
- Boroditsky, L. (2011). How language shapes thought. Scientific American, 304(2), 62–65.
- Brown, G. K., Beck, A. T., Steer, R. A., & Grisham, J. R. (2000). Risk factors for suicide in psychiatric outpatients: A 20-year prospective study. Journal of Consulting and Clinical Psychology, 68(3), 371–377.
- Chu, C., Buchman-Schmitt, J. M., Stanley, I. H., Hom, M. A., Tucker, R. P., Hagan, C. R., ... & Joiner, T. E. (2017). The interpersonal theory of suicide: A systematic review and meta-analysis of a decade of cross-national research. Psychological Bulletin, 143(12), 1313–1345.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
- Ishfaq, A., & Bhatti, A. M. (2019). Language shift and imminent language death: A diachronic study of Dawoodi. Harf-o-Sukhan, 3, 1–14.
- Ishfaq, A., & Bhatti, A. M. (2020). Lexical attrition and generational language competence in Dawoodi speakers. Harf-o-Sukhan, 4, 4–18.
- Ishfaq, A., & Bhatti, A. M. (2021). From pidgin to creole to collapse: The evolutionary trajectory of Dawoodi language. Jahan-e-Tahqeeq, 4.
- Ishfaq, A., Sultan, N., Hassan, N., Aleem, F., & Maldonado, M. G. (2022). Elif Shafak's Forty Rules of Love: Contextual variation in adjectives. International Online Journal of Language and Literature.
- Ishfaq, A., & Bhatti, A. M. (2022). Linguistic hegemony and the silencing of Dawoodi: Power, stigma, and structural marginalization. Jahan-e-Tahqeeq, 5(4), 48–60.
- Ishfaq, A., & Bhatti, A. M. (2023). Code-switching, borrowing, and linguistic dilution: Contact-induced change in Dawoodi. Jahan-e-Tahqeeq, 6(3), 577–592.
- Ishfaq, A., & Sultan, N. (2024). Identification of different methodologies for treatment of autism in Urdu-speaking adolescents: An investigative report. Contemporary Journal of Social Science Review, 2(4), 1611–1618.
- Ishfaq, A., & Bhatti, A. M. (2024). From heritage to liability: Language attitudes and identity reconstruction as drivers of obsolescence in the Dawoodi language. Al-Mahdi Research Journal (MRJ), 5(3), 1303–1336.
- Ishfaq, A., Malik, A. H., & Sultan, N. (2025). Developing trauma-sensitive pedagogical practices for resilient learning in academia: A multidisciplinary approach of psycholinguistics and ELT. Al Aasar, 2(1), 171–189.

- Ishfaq, A., Sultan, N., & Healy, B. (2025). Turn-taking, politeness, and identity: A conversational study of Speak Your Heart. Journal of Applied Linguistics and TESOL (JALT), 8(3), 1567–1581.
- Ishfaq, A., Azim, M. U. (2025). Phono-semantics and translation: A cross-linguistic study of Urdu and Punjabi ideophones. International Research Journal of Arts, Humanities and Social Sciences, 2(3).
- Ishfaq, A., Ahmad, S., & Sultan, N. (2025). Decoding despair: A multidisciplinary psycho-forensic linguistic approach to suicide notes. Journal of Psychology, Health and Social Challenges, 3(2), 83–89.
- Ishfaq, A., & Sultan, N. (2025). Narrative and meaning in Surah Yūsuf: A critical hermeneutic analysis. AL-HAYAT Research Journal (AHRJ), 2(4), 11–23.
- Ishfaq, A., & Sultan, N. (2025). Trauma, resilience, and narrative healing: A psycho-hermeneutic reading of Surah Yūsuf. AL-JAMEI Research Journal, 3(1), 229–239.
- Ishfaq, A. (2025). Ethnolinguistic identity and cultural memory in the Dawoodi community. Annual Methodological Archive Research Review, 3(6), 185–208.
- Khan, I. A., Khaled, F., & Ishfaq, A. (2025). Operation Bunyan un Marsoos: A critical analysis of human rights compliance—A study of the operation's adherence to human rights law and international humanitarian law. Dialogue Social Science Review (DSSR), 3(6), 173–184.
- Ishfaq, A., & Sultan, N. (2025). Cognitive control and executive function in advanced second-language writing. Sareer-a-Khama, 4(4).
- Sultan, N., & Ishfaq, A. (2026). Instagram captions as identity performance: A multimodal discourse analysis. AL-HAYAT Research Journal (AHRJ), 3(2), 9–21.
- Ishfaq, A., & Sultan, N. (2026). Modeling meaning in generative AI: A corpus-assisted discourse analysis of coherence, framing, and persuasion. Pakistan Journal of Social Science Review, 5(3), 1147–1164.
- Ahmad, A. I. S. (2026). Language of influence: A corpus-based lexical analysis of psychological power strategies in Robert Greene's The Laws of Human Nature. In Proceedings of the 2nd Riphah International Conference on Language, Literature, and Culture.
- Ishfaq, A. (2026). Thinking through language: Linguistic foundation and advanced academic writing
- Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, 51–60.
- Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology.
- Fawcett, J., Scheftner, W. A., Fogg, L., Clark, D. C., Young, M. A., Hedeker, D., & Gibbons, R. (1990). Time-related predictors of suicide in major affective disorder. American Journal of Psychiatry, 147(9), 1189–1194.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. Studies in Linguistic Analysis, 1–32. Blackwell.
- Fussell, S. R. (Ed.). (2002). The verbal communication of emotions: Interdisciplinary perspectives. Lawrence Erlbaum Associates.
- Gabrielatos, C., & Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Presented at the CADS International Conference, University of Bologna.
- Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... & Sheth, A. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. Proceedings of The Web Conference 2019.
- Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports, 7, 45141.
- Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.
- Ingram, R. E. (1990). Self-focused attention in clinical disorders: Review and a conceptual model. Psychological Bulletin, 107(2), 156–176.
- Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021). Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, 8(1), 214–226.
- Joiner, T. E. (2005). Why people die by suicide. Harvard University Press.
- Kövecses, Z. (2000). Metaphor and emotion: Language, culture, and body in human feeling. Cambridge University Press.
- Leenaars, A. A. (1988). Suicide notes. Human Sciences Press.

- Luoma, J. B., Martin, C. E., & Pearson, J. L. (2002). Contact with mental health and primary care providers before suicide: A review of the evidence. American Journal of Psychiatry, 159(6), 909–916.
- Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. Psychological Bulletin, 128(4), 638–662.
- O'Connor, R. C. (2011). Towards an integrated motivational–volitional model of suicidal behaviour. In R. C. O'Connor, S. Platt, & J. Gordon (Eds.), International handbook of suicide prevention (pp. 181–198). Wiley-Blackwell.
- O'Connor, R. C., & Kirtley, O. J. (2018). The integrated motivational–volitional model of suicidal behaviour. Philosophical Transactions of the Royal Society B, 373(1754).
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, 54(1), 547–577.
- Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., ... & Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights, 5, 3–16.
- Pirkis, J., & Burgess, P. (1998). Suicide and recency of health care contacts. British Journal of Psychiatry, 173(6), 462–474.
- Priel, B., Mitrany, D., & Shahar, G. (1998). Closeness, support and reciprocity: A study of attachment styles in adolescence. Personality and Individual Differences, 25(6), 1183–1197.
- Rayson, P. (2008). Wmatrix: A web-based corpus processing environment. Lancaster University.
- Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL Semantic Analysis System. Proceedings of the LREC Workshop on Beyond Named Entity Recognition.
- Scott, M. (1997). PC analysis of key words – and key key words. System, 25(2), 233–245.
- Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology.
- Shneidman, E. S. (1993). Suicide as psychache: A clinical approach to self-destructive behavior. Jason Aronson.
- Shneidman, E. S. (1996). The suicidal mind. Oxford University Press.
- Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.
- Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. Psychosomatic Medicine, 63(4), 517–522.
- Tay, D. (2012). Applying the notion of metaphor types to compare counseling and everyday talk. Journal of Counseling & Development, 90(3), 347–356.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.
- Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E. (2010). The interpersonal theory of suicide. Psychological Review, 117(2), 575–600.
- World Health Organization. (2023). Suicide worldwide in 2019: Global health estimates. WHO Press.
- Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology.