Hebrew Psychological Lexicons

Natalie Shapira, Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Adar Paz, Dana Stolowicz-Melman, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, Chen Dahbash, Limor Dayan, Tamar Naim, Lidar Gez, Boaz Yanai, Adva Maman, Adam Nadaf, Elinor Sarfati, Amna Baloum, Tal Naor, Ephraim Mosenkis, Matan Kenigsbuch, Badreya Sarsour, Yarden Elias, Liat Braun, Moria Rubin, Jany Gelfand Morgenshteyn, Noa Bergwerk, Noam Yosef, Sivan Peled, Coral Avigdor, Rahav Obercyger, Rachel Mann, Tomer Alper, Inbal Beka, Ori Shapira, Yoav Goldberg
Bar-Ilan University, Israel

Abstract

We introduce a large set of Hebrew lexicons pertaining to psychological aspects. These lexicons are useful for various psychology applications such as detecting emotional state, well being, relationship quality in conversation, identifying topics (e.g., family, work) and many more. We discuss the challenges in creating and validating lexicons in a new language, and highlight our methodological considerations in the data-driven lexicon construction process. Most of the lexicons are publicly available, which will facilitate further research on Hebrew clinical psychology text analysis. The lexicons were developed through data driven means, and verified by domain experts, clinical psychologists and psychology students, in a process of reconciliation with three judges. Development and verification relied on a dataset of a total of 872 psychotherapy session transcripts. We describe the construction process of each collection, the final resource and initial results of research studies employing this resource.

1 Introduction

A lexicon is the vocabulary of a domain of knowledge, and can be a valuable tool in the analysis of many psychological tasks. For example, in detecting clients' mental states, emotions and symptoms (Guntuku et al., 2017; Trotzek et al., 2018).

Lexicons are especially advantageous when data is scarce. Often in psychotherapy research, few samples are available in clinical trials, and confidentiality limits sharing of data. Scarcity of data is particularly challenging in less common languages like Hebrew. Recent data-hungry models are not practical in such cases where data is small, while other approaches, applying the use of lexicons, are more effective for predictive abilities. Moreover, lexicons can be shared across studies and serve as *clinical markers* (e.g., Al-Mosaiwi and Johnstone, 2018).

Additionally, through their simplicity, lexicons enable easy interpretation of results. They can be elaborate for indicating psychological states within text, e.g., in accordance to the frequency of specified terms within a passage (Tausczik and Pennebaker, 2010).

Lexicons are widely used in research and industry due to their proven effectiveness and ease of use. There are several psycho-linguistic lexicons, amongst them the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), Vader-lexicon (Hutto and Gilbert, 2014), NRC-Sentiment-Emotion-Lexicon (Mohammad and Turney, 2013), MRC (Coltheart, 1981), and DLATK (Schwartz et al., 2017), however *no valid psycho-linguistic lexicon for Hebrew exists*. [1]

Several approaches are generally employed for developing lexicons. One prevalent method involves judging collected words with domain experts (Pennebaker et al., 2015) or with crowdsourcing (Tanana et al., 2016). There are also various methods for translating existing lexicons from other languages (e.g., triangulation-based, machine translation and then manual fine-tuning). However lexicon translation tends to be impractical since direct translation leads to incomplete or wrong results (Massó et al., 2013) . In particular, the Hebrew language poses many word-level translation obstacles due to its morphologically-rich form and ambiguous orthography (as outlined in Section 2).

We describe the development of a collection of Hebrew psychological lexicons that were created between the years 2018 and 2021. We utilize a base dataset of 872 psychotherapy sessions, described in Section 3, to either validate or extract words for the lexicons. The first set of lexicon collections (Section 4) are devised by domain experts, and verified using the base dataset. The word lists in the second set (Section 5) are fully automatically generated

[1] A large collection of Hebrew NLP resources are available at https://github.com/NLPH/NLPH.

	Expert Knowledge Based Lexicons				Data-Driven Lists		Expert Knowledge + Automatic Methods	
					Supervised	Unsupervised	Translation	Expansion
Collection name	**Valence (Positive-Negative)**	**Emotional Variety**	**Paralinguistic**	**Depressive Characteristics**	**Well-Being**	**Conversation Topics**	**Hebrew LIWC**	**Extended Emotional Variety**
Number of lexicons/lists	2	42	11	14	2	200	∼40 out of 125	44
Total number of words	200	7313	154	194	40	4000	under construction	under construction
Coverage	2000 most frequent word types in dataset	5000 most frequent word types in dataset	31,067 tokens 1022 word types	several hundred most important word types	139 non-clinical sessions 38 clinical sessions	the whole dataset ∼5 million tokens	-	-
Verified by at least three domain experts	yes	yes	yes	yes	-	-	yes	under construction
Initial research use case	yes	work in progress	yes	yes	-	yes data-dependent	-	-
Freely available	yes	yes	yes	yes	yes	yes	internal use only	will be released

Table 1: A summary of the presented lexicons and word lists.

based on the dataset, and mainly serve for textual analysis of psychotherapy sessions. Section 6 combines domain experts and automatic methods for the preparation of lexicons. For each of the lexicon collections and methods, we provide a use-case in the clinical psychotherapy domain, illustrating their usefulness and effectiveness. See Table 1 for a description and statistics on the lexicons.

While many of the lexicon types described are common in the psychology domain, we additionally introduce two new lexicon types. The first is an *emotional-variety* lexicon type with *complementary-emotions*, i.e., each emotion lexicon has a complementing-emotion lexicon, valuable for reducing noise when analyzing emotion. The second type is for *paralinguistic* categorization, which enables the classification of different non-verbal vocal behavioral events within psychotherapy sessions.

Most of the lexicons freely available,[2] which will facilitate further research on Hebrew clinical psychology text analysis. The methods described may also aid in the establishment of additional lexicons in Hebrew and in other languages.

2 Challenges with Lexicon Translation

While methods for translating existing lexicons from other languages have been exploited before, lexicon translation yields wrong categorization of words (Massó et al., 2013). This is particularly the case when involving morphologically rich languages, and is also due to word ambiguity and cultural influence on languages.

In Hebrew, like in other Semitic (e.g., Arabic) and Indo-European languages (e.g., Spanish, Dutch), there are inflections and verb conjugations that have no direct conversion in English. Van Wissen and Boot (2017) address the problem by converting each word in a lexicon to its lemma (i.e., canonical form) and then using an existing list to expand to the various linguistic conjugations. In Hebrew it is possible to retrieve all the different inflections and verb conjugations for many words using specialized linguistic lexicons, such as the MILA lexicon (Itai and Wintner, 2008).[3] Even so, it is not always the case that all forms of a word should be included in the same lexicon. For example, in the *emotion variety* lexicon collection (Section 4.2), the word **רגוע** 'ragua' (relaxed) appears in the *not-nervous* lexicon and **תרגיע** 'targia' (calm down) appears in the *not-guilty* lexicon, sharing the same root form but having different semantic emotional classification.

In addition, there may be situations of ambiguity in which words with completely different meanings are mapped to the same lemma, e.g., the words (1) **חימה** 'chema' (anger) and **חמה** 'chama' (sun) have the same orthographic lemma **חמה**; (2) **עדשות** 'adashot' (contact lenses) and **עדשים** 'adashim' (lentils) have the same orthographic lemma **עדשה** 'adasha', thus adding noise to the directly-translated lexicon.

Furthermore, when expanding a lexicon around a word, ignoring diacritics often yields ambiguous forms. For example, while the word **אחלה** 'achla' (cool) is in the *positive emotion* lexicon (Section 4.1), without diacritics the optional base forms are **איחל** 'ichel' (wish), **חילה** 'chila' (to make ill), **אחלה** 'achla' (cool) and **חלה** 'chala' (to become ill), having different emotional polarity. Then, each of these words is also expanded with all their inflections, e.g., **חליתי** 'chaliti' (I became ill), adding up to hundreds of words to the wrong lexicon.

[2] https://github.com/natalieShapira/HebrewPsychologicalLexicons. As LIWC is commercial, we cannot publicly release the translated lexicons described in Section 6.1

[3] We use the BGU-version of the lexicon, which is bundled with the YAP Hebrew parser (More and Tsarfaty, 2016) as the file `bgulex.utf8.hr`.

Another problem is that there are lexicon types whose translation is not straightforward. For example, the *I words* lexicon in LIWC is a small set of 12 distinct words (e.g., *I, me, mine*) (Tausczik and Pennebaker, 2010) and can be used to count the frequency of all the occurrences of first-person mentions in a given text passage in English. However, Hebrew's morphological system preclude such word-counting method for seeking "I words" in the text passage, as the first-person status is often realized morphologically, and may appear on many word forms. Hebrew words follow a complex morphological structure, with both derivational and inflectional elements, that can encode gender, number, tense, person, possessive and noun-compounding. For example, אהבתי 'ahavti' (I loved), אוהב 'ohav' (I will love), אוהבת 'ohevet' (I-feminine love/she loves), אהובי 'ahuvi' (my love), Therefore, preprocessing of syntactic and morphological parsing is a critical phase for extracting the relevant details (e.g., the first person singular counts).

Lastly, the ambiguous interpretation in different languages makes out-of-context translation impossible. For example, the word 'dear' will be translated in Hebrew to the word יקר 'yakar' , but יקר 'yakar' also means 'expensive'. While 'dear' in LIWC is a word with positive polarity, 'expensive' is not. We cannot assume that if a resource is valid in language A, then its translation into language B will necessarily give us a valid resource in language B.

Relatedly, language is strongly culturally influenced, and a word may be categorized differently across languages and cultural context in terms of human psychology, especially around emotion or sentiment (Wierzbicka, 1985). For example, the color green, will refer to jealousy and envy in some cultures: "green-eyed monster" was first used by William Shakespeare about jealousy. There are proverbs in Hebrew that associate envy to the green color: "green with envy". In addition, in Hebrew ירוק ('yarok' green) can be used as a mockery of a person with no experience in his or her field, like an unripe fruit, especially used in the military context—a recruit. In contrast, green serves as a religious/sacred symbol in Islam as Muhammad's favorite color. (See also cultural differences in a study that examined the relationship between colors and emotions by Hupka et al., 1997.)

3 Base Dataset Description

All our lexicons rely on a dataset[4] of a total of 872 psychotherapy session transcripts from 74 different client-therapist dyads (pairs) consisting of a total of about 5 million tokens—100 thousand word types (unique words). All sessions are labeled with psychological analysis information that assists in generating a lexicon and/or verifying one. We infer relevant session-level labels from questionnaires filled by the participants at each session: (1) clients self-reported their well-being, measured using the ORS questionnaire (Miller et al., 2003), which is considered to be an indicator for progress in treatment; (2) therapists and clients reported on interpersonal relational events that occurred during a session, corresponding to tensions or breakdowns in their collaborative relationship (alliance ruptures), measured by the PSQ questionnaire (Muran et al., 2004); (3) therapists and clients reported emotional states measured by the POMS questionnaire (McNair, 1992).

4 Lexicons Based on Expert Knowledge

The approach employed for creating the following lexicons is inspired by that of Pennebaker et al. (2015), specifically via a three-judge (domain experts) reconciliation procedure for admitting words into a lexicon.

4.1 Valence (Positive and Negative)

A fundamental aspect to consider in psychological analysis is detecting positive and negative emotion. With regards to clinical text analysis, words identified as emotionally positive or negative have been shown to correlate to clinical conditions (Morales et al., 2017).

To create the positive and negative emotion lexicons, we collected the 2000 most frequent words (including stop words) from our base data as candidates. We found that these 2000 most frequent words cover 86% of all tokens in all transcripts. Three judges independently rated whether each word should be categorized as generally having a positive and/or negative emotion, after which a reconciliation process was conducted to resolve conflicting decisions. Initial Fleiss' Kappa (Fleiss, 1971) for interrater agreement was 0.54 (moderate

[4] See the appendix for more details about the participants, demographics information, treatment, transcriptions, questionnaires and ethical concerns.

agreement) and the final was 0.95, indicating almost perfect agreement (Landis and Koch, 1977). The main changes following the reconciliation process was (1) the addition of words with low polarity/confidence e.g., the word אבל 'aval' (but) was added in the second phase to the negative list; (2) the correction of errors and mistakes e.g., the word אוקי 'okay' (OK), was included in the positive list while the word אוקיי which is the same meaning 'okay' (OK), was not included; (3) better agreement on 'mixed emotion words' that evoked both positive *and* negative emotions (8.7% e.g., mother, feeling, power) compared to words evoking any emotion (73% e.g., also, like, type). There were no words with hard disagreement, i.e., where at least one of the judges marked the word as positive only and another judge marked it as negative only. In total, the lexicons contain 200 positive and negative emotion word types. To avoid ambiguities and encourage uniformity between future studies, we released only one version of lexicons (majority of two judges excluding mixed emotion words).[5]

Based on the two lexicons, we calculated the number of positive and negative emotion words within each session transcript (an hour of conversation) in the dataset. On average, there were 185 positive emotion words and 327 negative emotion words per session. 15% of the all tokens in the transcripts were emotion words.

Usage In one study conducted in our lab, we found correlations between a client's and therapist's positive/negative emotion words and client's and therapist's positive/negative emotions as reported in the POMS questionnaire. In another study, that uses our positive-negative emotion lexicons, Shapira et al. (2020) examined the relationship between the number of emotion words spoken in a session and the client's self-reported questionnaire regarding her well-being. The findings are consistent with the literature and in line with theoretical views highlighting the role of positive emotions and negative emotions and the association to well-being (e.g., Blatt (1995); Shahar et al. (2020); Morales et al. (2017)). Finally, Juravski (2020) also shows a correlation between the use of positive and negative emotion lexicons to predicted emojis by a pretrained model based on Twitter data,[6] contributing to the mutual validation of the tools. The above studies show that positive and negative emotion lexicons can be leveraged for automatic detection of emotional state and well-being within texts.

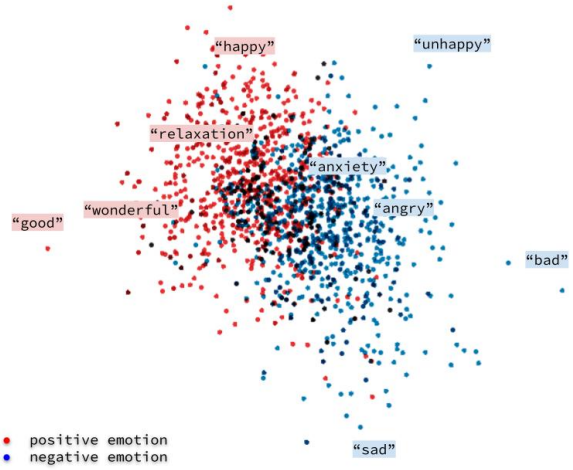

Figure 1: 2D-Projection of emotion word embeddings.

4.2 Emotional Variety

A great and diverse variety of emotional states exist, and in this section we describe the process of developing lexicons that relate to this variety. Our motive for developing these emotional lexicons stems from a basic notion in psychotherapy research: the ability to be in touch with emotional experiences, to portray them in words and to give them meaning, as a result of treatment, has been found to effectively predict improvement in mental well-being. This is consistent across various therapeutic models and types of mental disorders (Greenberg et al., 2012).

The development of the emotion lexicon was carried out in several stages. We first compiled a list of emotions on the basis of the POMS emotion questionnaire (see Appendix A.2.2), Robert plutchik's "wheel of emotions" (Plutchik, 2000) and those described by Ong et al. (2018). The list includes: *enthusiastic*, *amused*, *proud*, *interested*, *calm*, *sad*, *ashamed*, *guilty*, *hostile*, *nervous*, *anger*, *contentment*, *anxiety*, *vigor*, *joy*, *disgust*, *surprise*, *trust*, *anticipation*, *confusion*, *fatigue*.

For each emotion we created another category that is the complement of that emotion (e.g. *not_sad* as the complement of *sad*), hence resulting in a total of 42 categories.

The main purpose for categorizing complementing emotions is to enable more precise word categorization when requiring emotional analysis of

[5] Other versions (e.g. consensual words, words with low polarity, mixed emotions words) can be obtained upon request.

[6] https://hub.docker.com/r/danieljuravski/hemoji

text. An additional important motive is the long-term thought for allowing automatic expansion of these lexicon seeds (Section 6.2) using semantic-based methods.[7] Having a complementing-emotion word list can assist in the expansion process of the corresponding emotion lexicon by providing indicators for what might *not* categorize to that emotion. Figure 1 shows the projection of a list of positive and negative (complementing) emotion word embeddings.[8] While most words indeed separate to two different clusters, the clusters intersect considerably. This illustrates that it is not enough to assume that words will semantically cluster together by their emotional category. Having an emotion's complementary lexicon can be advantageous for finding new words for that emotion.[9] To the best pf our knowledge, we are the first to propose complementary-emotion lexicons.

In the second stage of the lexicons' development, 19 advanced undergraduate psychology students were given the list of emotional categories and were asked to suggest at least five appropriate words for each. Words could be produced either associatively or through active search (e.g., by using an online Hebrew thesaurus [10]). We additionally conducted a similar classification annotation procedure as described in Section 4.1, whereas in this case the 5000 most frequent words, covering 90% of all tokens in all transcripts, were tagged with one of the 42 emotion categories (not every word evoked an emotion). These were merged with the freely-suggested words from above.

The final collection of emotional variety lexicon seeds consists of a total of 7313 emotion words. The percentages of judges' agreement for the rating phase ranged from 98% to 100% agreement. This lexicon collection is available as a ready-to-use version. An expanded version of this lexicon is currently in the works (with the algorithm mentioned above, in Appendix A.3).

[7] Such as with the *word-similarity* package, pretrained on Hebrew Twitter word embeddings. https://github.com/Ronshm/hebrew-word2vec

[8] Using the *Tensorflow Embedding Projector* tool. https://projector.tensorflow.org

[9] See Appendix A.3 for a potential algorithm that could be used to expand emotion lexicons, using the complementing lexicon.

[10] such as https://synonyms.reverso.net/synonym/he/

Therapist: Shall I get you a glass of water? *<In a whisper>*

Client: *<Sounds of silent crying. Pulling the nose>* yes *<Like clearing throat>*, yes.

Figure 2: An example of paralinguistic event annotations (in *italics*) within the transcription, described in free text by the transcriber.

4.3 Paralinguistic Events

Paralinguistic events refer to non-verbal vocal elements of interpersonal language communication that accompany the verbal message. This component of communication may change meaning, create nuance or convey emotion, through the use of various techniques such as pitch and volume, weight, intonation, silences, laughter, etc. (Valstar et al., 2013), and may be expressed consciously or unconsciously (Harris and Rubinstein, 1975) by participants. Sometimes these elements are considered aphonemic, i.e., they cannot even be spelled out (Trager, 1961). All of these phenomena are inherent in the speech sequence, and are often processed as words in automatic speech processing – a *high tone* in speech as an indication of anxiety or a *breathy voice* as an indication of attractiveness – are already processed into the voice message.

Paralinguistic elements are of great importance in the therapeutic context. To date, much credible evidence has accumulated in research that confirms that characteristics of voice significantly influence the formation and development of the therapeutic relationship (Sikorski, 2012). In the clinical setting, paralinguistic communication is of fundamental importance to therapist-client dynamics. For example, through unconscious perception of change in the client's paralinguistic events, the therapist (while noticing the overt meaning conveyed through semantic channels) can adjust his or her own paralinguistics, and with a good understanding of the client's inner state, he or she can encourage expansion of the client's awareness (Rocco et al., 2013). Moreover, a strong association between vocal characteristics and certain psychopathological states has been documented, e.g., depression accompanied by slow, long, and intertwined speech in breaks (Ellgring and Scherer, 1996).

The paralinguistic events were labeled (as comments) in our transcripts dataset by the transcribers as free text (see examples in Figure 2). A total of 31,067 tokens occur in the transcriber comments, of which 2147 are unique and 1022 appear at least twice. The most frequent tokens are: "laughing"

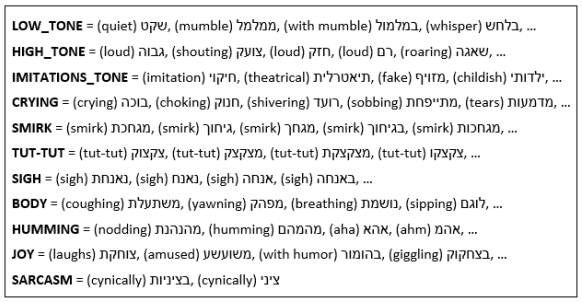

Figure 3: Paralinguistic categories (lexicons) and examples of words within them.

(feminine singular) at a frequency of 22%, "laughing" (masculine singular) at 5.3%, "tut-tut" (3.5%), "sigh" (2.5%), "laugh" (feminine plural; 2.3%), "giggle" (feminine singular; 1.8%), "of" (1.2%), "tongue" (1.2%), "cry" (referred to in masculine and feminine alike; 1.2%), "the therapist" (1%), "chuckle" (1%), "coughing" (1%), etc.

An NLP researcher, a clinical psychologist and two interning therapists went over the labels and their frequencies together and characterized 11 categories of paralinguistic events that are meaningful in psychological treatment: *low tone*, *high tone*, *imitation tone*, *crying*, *smirk*, *tut-tut*, *sigh*, *body-related*, *humming*, *joy*, and *sarcasm*. Then, each of the labels was classified into these categories (classification was trivial with 100% agreement, see Figure 3).

An initial study we conducted found strong correlations between paralinguistic events to postive and negative emotion words within psychotherapy sessions, e.g., strong positive correlation ($r=0.823$, $p < 0.001$) between *joy* paralinguistic events and positive emotion words within the therapist's text.

4.4 Depressive Characteristics

Depression is one of the most common mental disorders. In 2017, it was estimated that more than 300 million people worldwide (4.4% of the global population) were suffering of depression (WHO et al., 2017). Many studies have examined the relationship between depression and language (Trotzek et al., 2018; Yates et al., 2017; ODea et al., 2018; Ramirez-Esparza et al., 2008; Rude et al., 2004; Holtzman et al., 2017; Al-Mosaiwi and Johnstone, 2018; Ophir et al., 2020; Fineberg et al., 2016; Tackman et al., 2019; Guntuku et al., 2017; Morales et al., 2017; Tausczik and Pennebaker, 2010).

Referring to textual characteristics found in the above-mentioned literature, an NLP researcher and an interning therapist examined the sessions in the base dataset, and prepared a list of categories characterising depressive behavior, each category containing a list of characteristics. See Figure 4 for these characteristics.

Self-reference: first person singular, I words, changes belong to personal pronouns, possessive and pronouns based on POS tagging, Many third person pronouns, Unrelated personal pronouns ("it")
Emotions: Negative Emotions, Positive Emotions, Negative Content, Sadness, Anger, Anxiety, Negative attitude towards others compared to non-depressed with positive
Absoluteness spectrum: absolute, extreme, oath, hesitation, lack of fluency, tentative
Time and space: past, present, future, month of the writings, location
Text length: number of words, number of letters
Direct expression related to depression: "my depression", "my anxiety", "my therapist", "I was diagnosed with depression", Antidepressants e.g., "Zoloft", "Paxil"
Data-driven top phrases: "I went to", "my whole", "sometimes I", "I'm so sorry", "to scare you", "to have it", "my son was", "it wasn't"
Lyrical and abstract writing (life, time, values and religion) compared to non-depressed who are characterized by concrete prose writing (days, events, places, behaviors) and less reference to time
Miscellaneous: death related words, perceptual processes, article, contradiction (said, could have), attention to ingestion, curses, conditions ("if"), negation, interrupted and uncommitted, questions and question marks, necessity ("need") words compared to fewer words of desire ("love", "want"), swirls, not concrete (lots of words but little variety, short sentences, three points, fillers words as "like", unknown "don't know", shame, disappointment, repetitive, passive/active, numbers, helplessness, avoiding, repression, generalization (general talk and not about specific details), reputation, physical health, financial status, respect esteem, self-confidence

Figure 4: Linguistic characteristics of depressive texts, grouped by characteristic categories. We created lexicons for 14 of these characteristics.

Then, characteristic words were compiled in the following manner. A Random Forest classifier (Liaw and Wiener, 2002) was trained on all the clients' texts from the base data sessions, to predict the sadness-level label of a given text, as found in the POMS questionnaire of the corresponding session. A text was input to the classifier as a bag-of-words vector. Once the training completed, a few hundred of the most important features (words) were extracted from the trained classifier. These words were then categorized manually into 14 of the depressive characteristics, forming 14 new lexicons. One of these lexicons, for example, is called *tentativeness* (see under "Absoluteness spectrum" category in Figure 4), and consists of words such as כנראה (probably), אולי (maybe), and יתכן (perhaps). These word categorizations were then approved by two additional interning therapists.

5 Data-driven Word Lists

We next describe data-driven methods, applied on our base dataset, that extract lists of words for purposes of psychotherapetic analysis of session transcripts.

5.1 Well-Being

A potentially useful feature for automatically identifying outcome, i.e., improvement over psychotherapy treatment, is the client's well-being throughout

NON_CLINICAL_CONDITION = (punctuation) <PUNC>, (you) את, **(she) היא, (he) הוא**, (knows) יודעת, (xxx) XXX, **(him) לו, (her) לה**, (really) באמת, (with) עם, (I said) אמרתי, (ah) אה, (and) ו, **(her) אותה**, (also) גם, **(his) שלו**, (on) על, **(and she) והיא**, (always) תמיד, **(she was) הייתה**

CLINICAL_CONDITION = (but) אבל, (know) יודע, (then) אז, **(I) אני**, (such) כזה, (as) כאילו, **(that I) שאני**, (something) משהו, (it) זה, (yes) כן, (this) הזה, (say) נגיד, (which) איזה, (number) <NUM>, **(to me) לי**, **(I was) הייתי**, (em) אמ, (you) אתה, (can) יכול, (already) כבר

Figure 5: Data-driven lists of words characterizing clients in *non-clinical* condition versus *clinical* condition.

Topic 187	Topic 58	Topic 108	Topic 30	Topic 10	Topic 94	Topic 19	Topic 177
משפחה	עובד	בוקר	כסף	ללמוד	חרדה	מים	כלים
Family	Employee	Morning	Money	Learn	Anxiety	Water	Dishes
אמא	עבודה	לילה	לשלם	לימודים	שליטה	קפה	כביסה
Mother	Working	Night	Pay	Studies	Control	Coffee	Laundry
דודה	משרד	לישון	חשבון	תואר	פחד	כוס	מטבח
Aunt	Office	Sleep	Invoice	Degree	Fear	Glass	Kitchen
ילדים	אנשים	לקום	חודש	קורס	לשחרר	לשתות	מים
Children	People	Getting up	Month	Course	Release	Drink	Water
אחותי	מנהל	יום	בנק	אוניברסיטה	מובן	לקפוץ	מקלחת
Sister	Director	Day	Bank	University	Understandable	Jump	Shower
דודים	עסק	מיטה	מחיר	מבחן	זמן	יין	לשטוף
Uncles	Business	Bed	Price	Test	Time	Wine	Wash
אחים	בוס	שעה	דירה	תחום	עצבים	בקבוק	כיור
Brothers	Boss	Time	Apartment	Domain	Nerves	Bottle	Sink
סבתא	לקוחות	עייפה	עולה	מקצוע	גוף	בירה	מדיח
Grandmother	Customers	Tired	Costs	Profession	Body	Beer	Dishwasher
הורים	תחום	ללכת	סכום	שנה	התקף	שתייה	בגדים
Parents	Domain	Go	Amount	Year	Attack	Drink	Clothing
נכדים	שיווק	התעוררתי	משכורת	מתמטיקה	סטרס	קולה	מכונת כביסה
Grandchildren	Marketing	Woke	Salary	Math	Stress	Coca-Cola	Washing machine

Figure 6: A sample of topics.

the treatment. A collection of lexicons correlative to level of well-being (ranging from clinical, worst, to non-clinical condition, best) may assist in recognizing such patterns in treatment.

To extract data-driven lists of words that characterize client well-being, we followed the Marker Approach (Mergenthaler, 1996; Buchheim and Mergenthaler, 2000). First, the client texts from the base data sessions with the worst (0-8, clinical condition) and best (32-40, non-clinical condition) ORS questionnaire well-being scores were extracted. A total of 38 clinical and 139 non-clinical sessions were found in the data. Next, vocabularies were identified (Fertuck et al., 2012) for each of the two "worst" and "best" corpora in reference to each other. That is, words that are significantly more frequent in one text versus the other are marked. The top 20 words from each group was included in the final lexicons (see Figure 5). This set of lexicons did not go through an evaluation process yet.

Note that the emerging *clinical condition* lexicon includes words of first-person singular (FPS) form, which is consistent with the literature that finds an association between increased verbal use of the first-person and higher levels of distress (Tackman et al., 2019; Guntuku et al., 2017; Morales et al., 2017; Tausczik and Pennebaker, 2010). Moreover, this is in line with the theoretical literature that highlights the dominant role of self-focus and self-criticism in maintaining and intensifying individuals' negative affect, which in turn leads to increased symptoms of distress (Beck, 1967; Blatt, 1995; Pyszczynski and Greenberg, 1987; Shahar et al., 2020). Meanwhile, the *non-clinical condition* lexicon includes words of third-person singular (TPS), which might indicate a correlation to a healthier condition of well-being and speaking about others.

5.2 Conversation Topics in Psychotherapy

Therapists are driven to find methods for improving the quality of psychotherapy sessions, for example, by understanding whether the themes about which they converse with their clients influence the resulting outcome of the treatment. Hence, we wish to explore the topics within the sessions, and examine what words are characteristic of those topics.

We applied Latent Dirichlet Allocation (LDA; Blei et al. (2003)) on the transcripts data to detect clusters of words, occurring similarly within the psychotherapy sessions. This resulted in a set of 200 topics and their probability of appearing in the data (signifying how much weight they have in the psychotherapy data), with each topic containing a list of 20 words. Figure 6 shows a few examples of topics and their words, as generated from the data.

We find, for example, that topics 72, 15, 152, and 171 describe "celebration", "leisure experience", "enjoyment", and "choice", which intuitively seem to be related to positive experiences and to high functioning. On the other hand, topics such as 81, 199, 166, and 61 seem to be about "loneliness", "suffering", "physical difficulties", and "anger", which intuitively seem related to negative experiences and to low functioning.

We explored which topics (clusters) best identified clients' well-being and alliance ruptures (see Appendices A.2.1, A.2.4) and whether changes in these topics were associated with changes in outcome. A sparse multinomial logistic regression model was run to predict which topics best identified clients' functioning levels, and the occurrence of alliance ruptures in the sessions. Additionally, multi-level growth models were used to explore the associations between changes in topics and changes in outcome. The model identified the ruptures and outcome labels above chance (65%-75% accuracy). Change trajectories in topics were associated with change trajectories in outcome. The first four topics best correlated to a negative outcome. The results suggest that topic models can exploit rich linguistic data within sessions to identify psychotherapy

process and outcomes. For the detailed study see Atzil-Slonim et al. (2021).

It is important to note that the purpose of this section is to show a method for topic modeling, and not to produce topical-word lexicons for general use. The method should be reproduced on the data for which the analysis is required.

6 Lexicons Based on Expert Knowledge and Automatic Methods

This section describes lexicons that are *automatically* converted or expanded from existing *expert-based* lexicons.

6.1 Hebrew Translation for LIWC

Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) is the most famous lexicon collection in the field of psychological text analysis (tens of thousands of citations). LIWC contains 120 lexicons and has been incorporated in many research studies. A Hebrew translation of some of the LIWC lexicons, when possible, would contribute to aligned cross-lingual research. As LIWC is commercial, we cannot publicly release the translated lexicons described here, however the translation procedure we follow may be useful for other researchers seeking to translate certain lexicons.

Some of the categories are difficult or even impossible to translate into Hebrew. For example, the *articles* lexicon (e.g., "a", "an", "the", etc.) has no Hebrew equivalent,[11] nor does the *I words* lexicon (as explained in Section 2).

For lexicons that an equivalent can be produced (e.g. *family*, *work*, etc.), we suggest the translation process as follows: an LIWC lexicon contains a list of *prefixes* of words. In the first step, expand each prefix to all of its expanded forms using an English dictionary[12] (e.g., abandon* to: abandon, abandoned, abandoning, abandonment etc.). This provides a list of concrete words under each category (lexicon) instead of prefixes. In the second step, generate a list of optional translated words by translating each word via the word2word package[13] (Choe et al., 2019). This package provides 20 candidate translations for each word, hence each Hebrew-translated lexicon is 20 times the size of the respective English-LIWC lexicon. A total of about 150,000 words emerged for the translated lexicons. This number of words can be verified in about 1,000 hours by a three-judge verification process (estimating 500 words per judge per hour), which we are in the process of doing.

6.2 Expansions

As future work we plan to expand expert-knowledge-based lexicons, such as the *emotional variety* lexicon (Section 4.2), using automated methods. For example, we can automatically expand words on their inflection types, or find semantically similar words with, e.g., embedding-based expansions (for initial algorithm see Appendix A.3). Needless to say, the products of these methods will require expert validation procedures.

7 Limitations

The lexicons presented are based on a unique dataset of psychotherapy session transcripts. The language used by clients and therapists in these sessions do not necessarily reflect the language naturally occurring in other settings. Additionally, the statistical demographics of the participants in the utilized sessions are not fully balanced in terms of gender, age, education and relationship status (see Appendix A.1.1 for details). Again, this may influence the overall language observed, and in turn, the computations performed throughout our work in generating and verifying the lexicons.

8 Conclusion

We present a collection of novel Hebrew lexicons, based on psychological data and domain expert knowledge. We describe a variety of lexicon development methods: expert-knowledge-based, data-driven using labeled data and unsupervised learning. We address levels of reliability—agreement between three judges (expert knowledge) versus automatic methods that are vulnerable to noise. We describe the importance of the lexicons for psychology research, as well as initial uses cases with results.

The lexicons are released for the benefit of the community, contributing to psychological text-analysis research in Hebrew and cross-lingual research in general. Furthermore, we hope that the methods described will inspire the creation of additional lexicons in Hebrew and in other languages.

[11] The indefinite articles do not exist, while the definite article *the* is realized morphologically as a possibly ambiguous prefix which is attached to the token.

[12] E.g., the dictionary in SpaCy (Honnibal and Montani, 2017) or NLTK (Loper and Bird, 2002).

[13] Bilingual lexicons for 3,564 language pairs https://github.com/kakaobrain/word2word

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments and suggestions. This project has received funding from the Israel Science Foundation (grants 1348/15 and 1278/16); and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 802774 (iEXTRACT).

References

Mohammed Al-Mosaiwi and Tom Johnstone. 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.

Aviad Albert, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2013. The hebrew childes corpus: transcription and morphological analysis. *Language resources and evaluation*, 47(4):973–1005.

Dana Atzil-Slonim, Daniel Juravski, Eran Bar-Kalifa, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Natalie Shapira, and Yoav Goldberg. 2021. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*.

Aaron T Beck. 1967. *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press.

Matthew D Blagys and Mark J Hilsenroth. 2000. Distinctive features of short-term psychodynamic-interpersonal psychotherapy: A review of the comparative psychotherapy process literature. *Clinical psychology: Science and practice*, 7(2):167–188.

Sidney J Blatt. 1995. The destructiveness of perfectionism: Implications for the treatment of depression. *American psychologist*, 50(12):1003.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Anna Buchheim and Erhard Mergenthaler. 2000. The relationship among attachment representation, emotion-abstraction patterns, and narrative style: A computer-based text analysis of the adult attachment interview. *Psychotherapy Research*, 10(4):390–407.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2019. word2word: A collection of bilingual lexicons for 3,564 language pairs. *arXiv preprint arXiv:1911.12019*.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

James A Cranford, Patrick E Shrout, Masumi Iida, Eshkol Rafaeli, Tiffany Yip, and Niall Bolger. 2006. A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32(7):917–929.

Heiner Ellgring and Klaus R Scherer. 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110.

Fredrik Falkenström, Robert L Hatcher, Tommy Skjulsvik, Mattias Holmqvist Larsson, and Rolf Holmqvist. 2015. Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. *Psychological Assessment*, 27(1):169.

Eric A Fertuck, Erhard Mergenthaler, Mary Target, Kenneth N Levy, and John F Clarkin. 2012. Development and criterion validity of a computerized text analysis measure of reflective functioning. *Psychotherapy Research*, 22(3):298–305.

SK Fineberg, J Leavitt, S Deutsch-Link, S Dealy, CD Landry, K Pirruccio, S Shea, S Trent, G Cecchi, and PR Corlett. 2016. Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*, 46(12):2605.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Peter L Greenberg, Heinz Tuechler, Julie Schanz, Guillermo Sanz, Guillermo Garcia-Manero, Francesc Solé, John M Bennett, David Bowen, Pierre Fenaux, Francois Dreyfus, et al. 2012. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood*, 120(12):2454–2465.

Edward Guadagnoli and Vincent Mor. 1989. Measuring cancer patients' affect: Revision and psychometric properties of the profile of mood states (poms). *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 1(2):150.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Richard M Harris and David Rubinstein. 1975. Paralanguage, communication, and cognition. *Organization of behavior in face-to-face interaction*, pages 251–276.

Nicholas S Holtzman et al. 2017. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68:63–68.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Adam O Horvath and Leslie S Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Ralph B Hupka, Zbigniew Zaleski, Jurgen Otto, Lucy Reidl, and Nadia V Tarabrina. 1997. The colors of anger, envy, fear, and jealousy: A cross-cultural study. *Journal of cross-cultural psychology*, 28(2):156–171.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Alon Itai and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Daniel Juravski. 2020. Natural language processing methods for analysing textual psychotherapy data.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomforest. *R News*, 2(3):18–22.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Guillem Massó, Patrik Lambert, Carlos Rodríguez Penagos, and Roser Saurí. 2013. Generating new liwc dictionaries by triangulation. In *Asia Information Retrieval Symposium*, pages 263–271. Springer.

Douglas M McNair. 1992. Profile of mood states. *Educational and Industrial Testing Service*.

Erhard Mergenthaler. 1996. Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of consulting and clinical psychology*, 64(6):1306.

Erhard Mergenthaler and Charles Stinson. 1992. Psychotherapy transcription standards. *Psychotherapy research*, 2(2):125–142.

Scott D Miller, BL Duncan, J Brown, JA Sparks, and DA Claud. 2003. The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of brief Therapy*, 2(2):91–100.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology-From Linguistic Signal to Clinical Reality*, pages 1–12.

Amir More and Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016*.

J Christopher Muran, Jeremy D Safran, Bernard S Gorman, Lisa Wallner Samstag, Catherine Eubanks-Carter, and Arnold Winston. 2009. The relationship of early alliance ruptures and their resolution to process and outcome in three time-limited psychotherapies for personality disorders. *Psychotherapy: Theory, Research, Practice, Training*, 46(2):233.

JC Muran, JD Safran, LW Samstag, and A Winston. 2004. Patient and therapist postsession questionnaires, version 2004. *New York: Beth Israel Medical Center*.

Bridianne ODea, Tjeerd W Boonstra, Mark E Larsen, Thin Nguyen, Svetha Venkatesh, and Helen Christensen. 2018. The relationship between linguistic expression and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study of blog content. *arXiv preprint arXiv:1811.02750*.

Anthony D Ong, Lizbeth Benson, Alex J Zautra, and Nilam Ram. 2018. Emodiversity and biomarkers of inflammation. *Emotion*, 18(1):3.

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1):1–10.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Robert Plutchik. 2000. Emotions in the practice of psychotherapy-clinical implications of affect theories.

Tom Pyszczynski and Jeff Greenberg. 1987. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, 102(1):122.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Diego Rocco, Rachele Mariani, and Diego Zanelli. 2013. The role of non-verbal interaction in a short-term psychotherapy: Preliminary analysis and assessment of paralinguistic aspects. *Research in Psychotherapy: Psychopathology, Process and Outcome*, pages 54–64.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, pages 55–60.

Golan Shahar, Megan L Rogers, Hadar Shalev, and Thomas E Joiner. 2020. Self-criticism, interpersonal conditions, and biosystemic inflammation in suicidal thoughts and behaviors within mood disorders: A bio-cognitive-interpersonal hypothesis. *Journal of personality*, 88(1):133–145.

Natalie Shapira, Gal Lazarus, Yoav Goldberg, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Daniel Juravski, and Dana Atzil-Slonim. 2020. Using computerized text analysis to examine associations between linguistic features and clients' distress during psychotherapy. *Journal of counseling psychology*.

Jonathan Shedler. 2010. The efficacy of psychodynamic psychotherapy. *American psychologist*, 65(2):98.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, and Geoffrey C Dunbar. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *The Journal of clinical psychiatry*.

Wiesław Sikorski. 2012. Paralinguistic communication in the therapeutic relationship. *Arch Psychiatry Psychother*, 1:49–54.

Richard F Summers and Jacques P Barber. 2009. *Psychodynamic therapy: A guide to evidence-based practice*. Guilford Press.

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116(5):817.

Michael Tanana, Aaron Dembe, Christina S Soma, Zac Imel, David Atkins, and Vivek Srikumar. 2016. Is sentiment in movies the same as sentiment in psychotherapy? comparisons using a new psychotherapy sentiment database. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 33–41.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

George L Trager. 1961. The typology of paralanguage. *Anthropological Linguistics*, pages 17–21.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

Leon Van Wissen and Peter Boot. 2017. An electronic translation of the liwc dictionary into dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 703–715. Lexical Computing.

World Health Organization WHO et al. 2017. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Anna Wierzbicka. 1985. Different cultures, different languages, different speech acts: Polish vs. english. *Journal of pragmatics*, 9(2-3):145–178.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

A Appendices

A.1 Base Dataset Description

A.1.1 Clients

The dataset was drawn as a sample from a broader pool of clients who received individual psychotherapy at a university training outpatient clinic, located in a central city in Israel. Data were collected naturalistically between August 2014 and August 2016 as part of the clinic's regular practice of monitoring clients' progress. From an initial sample of 180 clients who provided their consent to participate in the study, 34 (18.88%) dropped out (deciding one-sidedly to end treatment before the planned termination date). Clients were selected from the larger sample to match two criteria: (1) treatment duration of at least 15 sessions, and (2) full data including audio recordings to be used for the transcriptions and session-by-session questionnaires available for each client. These criteria corresponded to our analytic strategy of detecting within-client associations between linguistic features and session processes and outcomes. Clients were also excluded, based on the M.I.N.I. 6.0 (Sheehan et al., 1998) if they were diagnosed as severely disturbed, either due to a current crisis, had severe trauma and accompanying post- traumatic stress disorder, a past or present psychotic or manic diagnosis, and/or current substance abuse. Based on these criteria we excluded 77 (42.7%) clients. Thus, of the total sample, the data for 68 (38.33%) clients who met the above-mentioned inclusion criteria were transcribed, for a total of 872 transcribed sessions.

The clients were all above the age of 18 (M_{age}=39.06, SD=13.67, range=20–77), majority of whom were women (58.9%). Of the clients, 53.5% had at least a bachelor's degree, 53.5% reported being single, 8.9% were in a committed relationship, 23.2% were married and 14.2% were divorced or widowed. Clients' diagnoses were established based on the Mini International Neuropsychiatric Diagnostic Interview for Axis I DSM-IV diagnoses (MINI 5.0; Sheehan et al., 1998). Of the entire sample, 22.9% of the clients had a single diagnosis, 20.0% had two diagnoses, and 25.7% had three or more diagnoses. The most common diagnoses were comorbid anxiety and affective disorders[14] (25.7%), followed by other comorbid disorders (17.1%), anxiety disorders (14.3%), and affective disorders (5.7%). A sizable group of clients (31.4%) reported experiencing relationship concerns, academic/occupational stress, or other problems but did not meet criteria for any Axis I diagnosis.

A.1.2 Therapists and Therapy

Clients were treated by 59 therapists in various stages of their clinical training. Clients were assigned to therapists in an ecologically valid manner based on real-world issues, such as therapist availability and caseload. Most therapists treated one client each (47 therapists), but some (10) treated two clients and (2) more. Each therapist received one hour of individual supervision every two weeks and four hours of group supervision on a weekly basis. All therapy sessions were audiotaped for supervision. Supervisors were senior clinicians. Individual and group supervision focused heavily on reviewing audiotaped case material and technical interventions designed to facilitate the appropriate use of therapist interventions. Individual psychotherapy consisted of once- or twice-weekly sessions. The language of therapy was Modern Hebrew (MH). The dominant approach in the clinic includes a short-term psychodynamic psychotherapy treatment model (e.g.,Blagys and Hilsenroth,2000; Shedler, 2010; Summers and Barber, 2009). The key features of the model include: (a) a focus on affect and the experience and expression of emotions, (b) exploration of attempts to avoid distressing thoughts and feelings, (c) identification of recurring themes and patterns, (d) an emphasis on past experiences, (e) a focus on interpersonal experiences, (f) an emphasis on the therapeutic relationship, and (g) exploration of wishes, dreams, or fantasies (Shedler, 2010). On average, treatment length was 37 sessions (SD = 23.99, range = 18–157). Treatment was open- ended in length, but given that psychotherapy was provided by clinical trainees at a university-based outpatient community clinic, the treatment duration was often restricted to be 9 months.

A.1.3 Transcriptions

To capture the treatment processes from session to session, and since the transcription process is highly expensive, transcriptions were conducted alternately (i.e., sessions 2, 4, 6, 8 and so on until

[14] The following DSM-IV diagnoses were assessed in the affective disorders cluster: major depressive disorder, dysthymia and bipolar disorder. The following DSM-IV diagnoses were assumed in the anxiety disorders cluster: panic disorder, agoraphobia, generalized anxiety disorder and social anxiety disorder.

one session before the last session). In cases where material was incomplete (such as the quality of the recordings, or the questionnaires for a specific session), the next session was transcribed instead. The transcriber team was composed of seven transcribers, all of whom were graduate students in the University's psychology department. The transcribers went through a one day training workshop and monthly meetings were held throughout the transcription process to supervise the quality of their work. The training included specific guidelines on how to handle confidential and sensitive information and the transcribers were instructed to replace names and places by pseudonyms and to substitute any other identifying information. The transcription protocol followed general guidelines, as described in (Mergenthaler and Stinson, 1992), and in Albert et al. (2013). The word forms, the form of commentaries, and the use of punctuation were kept as close as possible to the speech presentation. Everything was transcribed, including word fragments as well as syllables or fillers (such as "ums", "ahs", "uh huhs" and "you know"). The audiotape was transcribed in its entirety and provided a verbatim account of the session. The transcripts included elisions, mispronunciations, slang, grammatical errors, non-verbal sounds (e.g., laughs, cry, sighs), and background noises. The transcription rules were limited in number and simple (for example, each client and therapist utterances should be on a separate line ;each line begins with the specification of the speaker) and the format used several symbols to indicate comments (such as [...] to indicate the correct form when the actual utterance was mispronounced, or <number of minutes of silence >). The transcripts were proofread by the research coordinator. The final transcripts could be processed by human experts or automatically by computer.

There were 872 transcripts in total (the mean transcribed sessions per client was 12.56; SD=4.93) Each transcript incorporated metadata such as the client's code, which allowed the client data to be linked across sessions and for hierarchical analysis. The transcriptions totaled about four million words over 150,000 talk turns (i.e., switching between speakers). On average, there were 5800 words in a session, of which 4538 (78%; SD=1409.62; range 416-8176) were client utterances and 1266 (22%; SD=674.99; range 160-6048) were therapist utterances with a mean of 180.07 (SD=95.37; range 30-845) talk turns per session.

A.1.4 Procedure and Ethical Considerations

The procedures were part of the routine assessment and monitoring process in the clinic. All research materials were collected after securing the approval of the authors' university ethics committee. Only clients that gave their consent to participate were included in the study. Clients were told that they could choose to terminate their participation in the study at any time without jeopardizing treatment. The clients completed the ORS before each therapy session and the WAI after each session. The therapist completed the WAI after each therapy session. The sessions were audiotaped and transcribed according to a protocol described above. All data collected was anonymized (see Section A.1.3) and only then exposed to a very small number of researchers, as agreed upon by the participants. The data is stored encrypted.

A.2 Outcome and Process Measurements

A.2.1 Outcome Rating Scale (ORS; (Miller et al., 2003))

The ORS is a 4-item visual analog scale developed as a brief alternative to the OQ-45. The scale is designed to assess change in three areas of client functioning that are widely considered to be valid indicators of progress in treatment: functioning, interpersonal relationships, and social role performance. Respondents complete the ORS by rating four statements on a visual analog scale anchored at one end by the word Low and at the other end by the word High. This scale yields four separate scores between 0 and 10 that sum to one score ranging from 0 to 40, with higher scores indicating better functioning. The ORS has strong reliability estimates (α=0.87-0.96) and moderate correlations between the ORS items and the OQ-45 subscale and total scores (ORS total - OQ-45 total: r = 0.59).

A.2.2 Profile of Mood States (POMS; (McNair, 1992))

The POMS assesses mood variables and is widely used. For the purpose of this study, we used an abbreviated version of the measure, which was adapted for intensive repeated measurements (Cranford et al., 2006) and consists of 12 words that describe current emotional states. The negative affect scale includes depressed mood (2 items), anxious mood (2 items), and anger (2 items). The positive affect scale includes contentment (2 items), vigor

(2 items), and calmness (2 items). Examples of feelings on the POMS are 'anxious', 'sad', 'angry', 'happy', 'lively', and 'calm'. Clients were asked to evaluate how they felt during the session on a 5-point Likert scale ranging from 'Not at all' to 'Extremely'. The POMS has been tested on college students and was found to be both valid and reliable (Guadagnoli and Mor, 1989).

A.2.3 Working Alliance Inventory (WAI; (Horvath and Greenberg, 1989))

The WAI is a self report questionnaire (both for therapist and client). It is one of the most widely investigated common factors that was found positively correlated to treatment outcome in psychotherapy. It includes items ranging from 0 ("not at all") to 5 ("completely") to evaluate three components (1) agreement on treatment goals (2) agreement on therapeutic tasks and (3) a positive emotional bond between client and therapist (Falkenström et al., 2015)

A.2.4 Post-Session Questionnaire (PSQ; (Muran et al., 2004))

Alliance ruptures were assessed after each session with a single-item question from the therapist's perspective: "Did you experience any tension, any misunderstanding, conflict or disagreement in the relationship with your patient?" Both items are answered on a 5-point Likert scale ranging from 1 ("not at all") to 5 ("constantly"), reflecting the subjectively perceived intensity of a rupture. Following the recommendations provided by (Muran et al., 2009), a rupture was defined as any rating higher than 1 on the scale.

A.3 Expansion of Complementing Word Sets

This section formally defines the problem of expanding complimentary lexicons and describes technique as a solution.

Given:

1. $positive\_seed$, $negative\_seed$ which are two complementing lexicon seeds. E.g., Enthusiastic=[mighty, wow, energetic, ...] and the compliment Not_Enthusiastic=[apathetic, oh, nothing, ...]
2. $confidence\_level$, float greater than 0
3. $expand\_rate$, integer greater than 0
4. $radius$, integer greater than 1

Output:

$positive\_expansion$, $negative\_expansion$, new lexicons, each containing the given respective lexicon and additional words that match the lexicon's semantic knowledge.

Algorithm Intuition

The expansion is performed in several rounds, where in each round the two seeds $positive\_seed$, $negative\_seed$ expand simultaneously on the basis of words semantically similar to words that already exist in the seed. The generation process of new semantically similar words candidates uses the *word-similarity* package[15] that is based on pretrained Hebrew Twitter word embeddings, and returns similar words for a given word, with similarity probabilities. The $expand\_rate$ parameter represents the number of similar words that the *word-similarity* returns (default configured as 30).

While expanding, care is taken not to deviate from the lexicon to its complementing lexicon (to get a feel for the importance of this step, see Figure 1 of positive and negative emotion words, showing how semantically close the words in the complementing lexicons can be). Each word in the seed list is used as a "witness" for similar words (weighted by similarity probability). In case there is more than one "witness" for a new candidate word, the similarity probabilities are summed. This "sieve" process is done by making sure that for each word that enters the expansion lexicon there are enough "witnesses", other close words already in the existing seed lexicon (i.e., their sum of probabilities for similarity to the candidate word is above threshold for filter criterion) and also does not appear in the complementary lexicon. The $confidence\_level$ parameter (default configured to 3) represents the threshold for filter criterion.

The result of the expansion is used as input for the next round. The $radius$ parameter represents the number of expansion rounds.

Algorithm Steps

1. For $radius$ times:
   (a) For each of $positive\_seed$ and $negative\_seed$ seeds, create new sets of candidate words $positive\_candidates$ and $negative\_candidates$, by expanding the words in the seeds with

[15] https://github.com/Ronshm/hebrew-word2vec

word-similarity with $expand\_rate$ parameter as number of similar words.

(b) Each of $positive\_candidates$, $negative\_candidates$ passes a *candidates-sieve* process which creates $positive\_survivors$, $negative\_survivors$: filter out low-probability words (sum of probabilities less than $confidence\_level$) or words that appear in the complementary seed list (i.e., $negative\_candidates$ for the $positive\_candidates$ and vise versa) .

(c) Update seed lists $positive\_seed$ and $negative\_seed$ with the corresponding lists $positive\_survivors$ and $negative\_candidates$.

2. return $positive\_seed$, $negative\_seed$