A Appendices

A.1 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[15] (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

[15] 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 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 and only then exposed to a very small number of researchers, as agreed upon by the participants.

A.1.5 Missing Data

In the concurrent session-level models, from the transcribed sessions (872), 860 had functioning (ORS), 831 had therapist's therapeutic alliance (T_WAI) and 823 had client's therapeutic alliance (C_WAI). One transcription was detected with errors. Sessions with missing or faulty data were excluded from the analysis.

A.2 Outcome & 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 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.3 Complementing Behavior as Synchrony

Synchrony may be observed through complementing behavior, where the actions of one party influences the second party in a complementing manner, e.g., if a rise of an occurrence of a feature in the first party directly causes a proportional decline for the second party, and vice-versa, yielding a negative correlation.

We show here that the number of words spoken by the participants in the sessions renders such behavior. As one participant talks more within a session, the other naturally talks less. Since all psychotherapy sessions have a fixed length of one hour, we can comparably measure this effect across all sessions.

Algorithm 2: Client's (c) and therapist's (t) word count in sessions (size=m) correlation

candidateMLS-2(c,t,m);
for $j \leftarrow 1$ *to* m **do**
  $cWC_j \leftarrow wordCount(c_j)$;
  $tWC_j \leftarrow wordCount(t_j)$;
end
return: $pearsonr(cWC, tWC)$

We propose MLS function *CandidateMLS-2* (Algorithm 2) which receives as input lists C^d and T^d of size m_d of a client's and the matching therapist's transcribed speech within each of their sessions (m_d is the number of sessions within a specific dyad, d). Each list element contains the clients'/therapists' utterances from a single session, and $c_j^d \in C^d$ and $t_j^d \in T^d$ are from the same session, for each session j. The algorithm converts each element in the lists to the word-count-number. Finally, the algorithm outputs the Pearson coefficient correlation between the new lists.

A surrogate test (as describe in Section 5.3) produces significant separation both at the between-surrogate ($p < 0.05$ with large effect size, Cohen's d = 0.953) and within-surrogate ($p < 0.05$ with large effect size, Cohen's d = 1.038). These results shows that *CandidateMLS-2* is indeed MLS, notably featuring *complementing* synchrony.

A.4 LSM vs. POS

The LSM method (Ireland and Pennebaker, 2010) takes advantage of word categories defined in LIWC, see Table 3. LIWC was not translated to a Hebrew version. Languages behave differently and it is therefore impossible to produce a perfect translation. For example, in Hebrew there is no use of articles (for the challenges in the Hebrew translation process see Shapira et al., 2021).

Since a Hebrew LIWC version is not available, an alternative approach is to apply *part-of-speech* categories that can be loosely mapped to LIWC categories used in the LSM method. Part-of-speech (POS Marcinkiewicz, 1994) is a linguistic category of words that have similar grammatical properties, i.e., words assigned with the same part-of-speech tag play a similar role within the grammatical structure of sentences (for the multilingual efforts to create a universal POS tagset see Petrov et al., 2011).[16] The POS categories can express the way things are said rather than the content itself ("how" versus "what"). Extraction of POS tags is a common procedure in natural language processing, and relevant tools exist in Hebrew (e.g., YAP; More and Tsarfaty, 2016, see Table 4).

There is a loose relationship between LIWC categories used by LSM and the POS categories.

- The **Auxiliary** category in LIWC contains the words that fall under the COP POS category, but COP represents any copula (אוגד) which is not always a verb in Hebrew. In addition there is an intersection with the MD POS category (e.g., could).
- The **Conjunction** LIWC category can be mapped to the POS categories CONJ, CC, TEMP and REL. CONJ is for the coordinating conjunction ו (and); TEMP is for the subordinating conjunctions that precede time clauses e.g., כש (when); REL is for the relative clauses ה, ש (that); CC is for the rest of conjunctions, both coordinating and subordinating.

[16] For the universal POS tags see https://universaldependencies.org/u/pos/

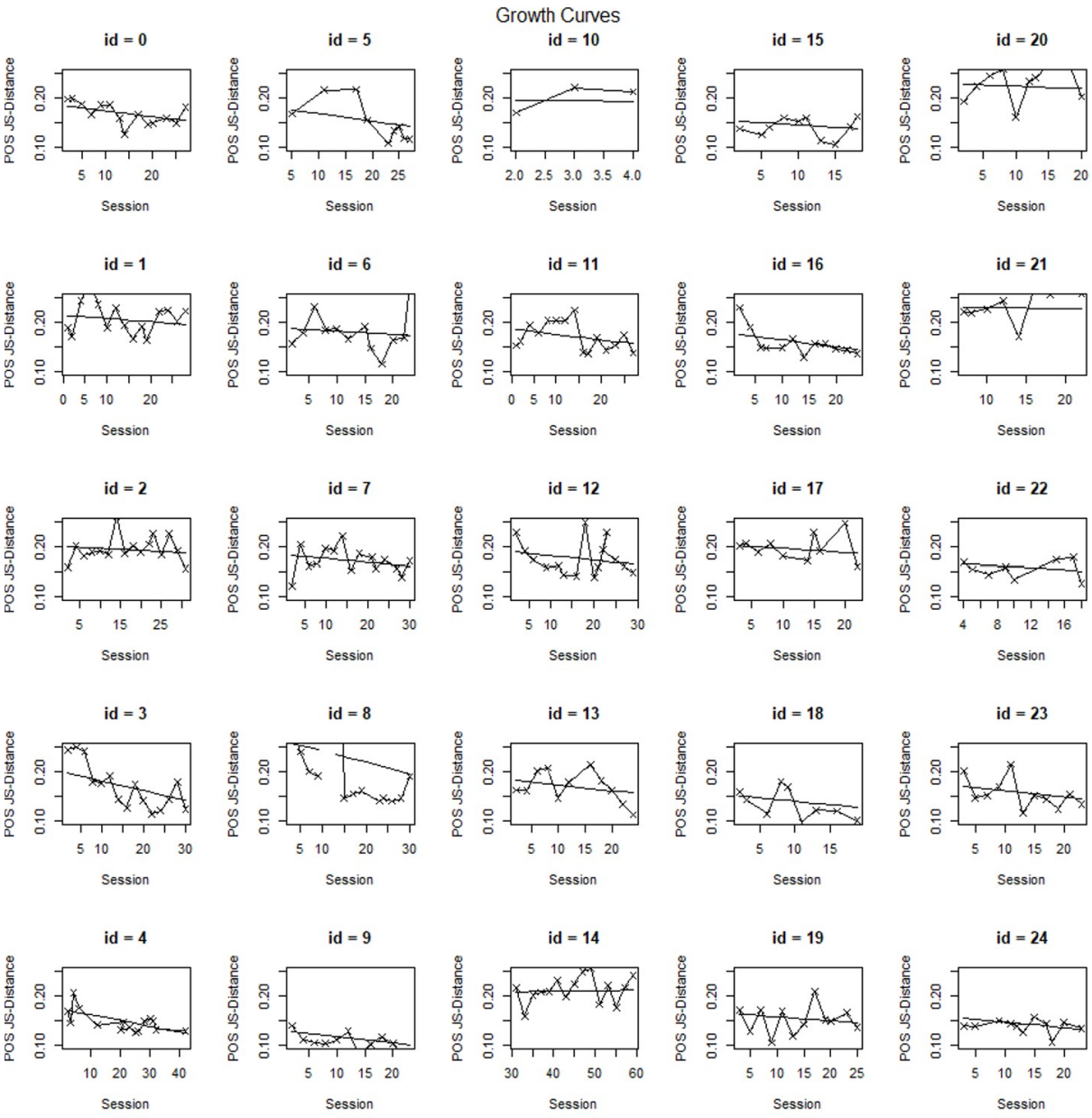

Figure 5: Growth Curves of 25 sampled dyads of the 74 available. There is a decrease of 0.001 units (i.e., slope) of JS-Distance between Probability Distribution over Unigram POS-tag in each session throughout treatment, indicating an increase in linguistic similarity. Results are statistically significant with $p<0.0001$.

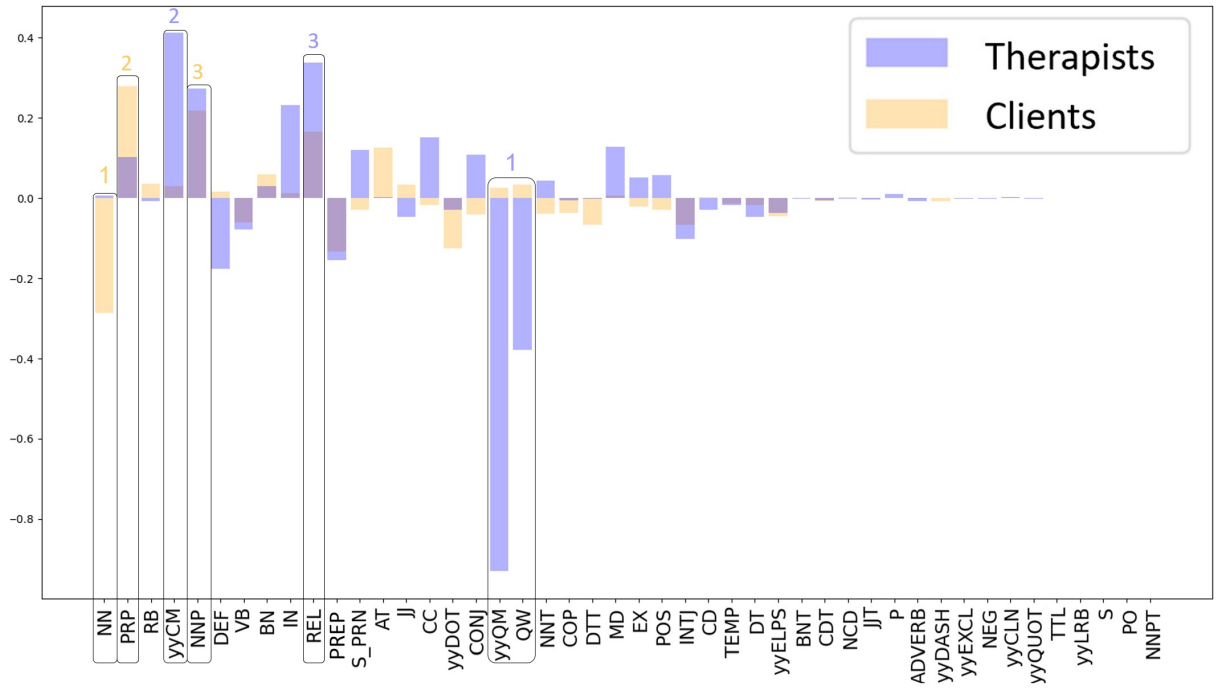

Figure 6: The sum of POS tag frequency changes between consecutive sessions for all clients (orange) and therapists (purple). A positive (negative) value means an overall increase (decrease) in the frequency of a POS tag throughout treatments. The three major changes in treatments for therapists are (1) decrease in questions (yyQM, QM) while for clients this increases, (2) increase in commas i.e., short break (yyCM), similarly to clients, (3) increase in “that" (REL), also similar to clients’ behavior. The three for clients are: (1) decrease in nouns (NN) while for therapists this increases, (2) increase in personal pronouns (PRP), as for therapists, and (3) increase in names (NNP) like for therapists. Overall, the therapists change throughout treatment more than the clients do.

- There is no POS category for the LIWC category **High-Frequency Adverbs**, but there is a POS category, RB, for general adverbs.
- The POS category PRP intersects with the LIWC categories **Personal** and **Impersonal Pronouns**. The POS category S_PRN is fully contained in the LIWC category of **Personal Pronouns** but only for single first person.
- The LIWC category **Negations** is partially represented by the POS category NEG.
- **Prepositions** with the LIWC categories can be mapped to the POS categories PREPOSITION and IN.
- **Quantifiers** with the LIWC categories can be mapped to the POS categories DT and DTT.
- In Hebrew there is no use of **Articles**.

In our study we used all possible POS categories.

LIWC LSM Categories

Category	Examples of Words in Lexicon
Articles	a, an, the
Auxiliary Verbs	ain't, am, are, ...
Conjunctions	also, and, as, but, ...
High-Frequency Adverbs	about, absolutely, actually, again, ...
Impersonal Pronouns	another, anybody, if, itself, ...
Personal Pronouns	he, him, ...
Prepositions	about, above, along, ...
Quantifiers	add, alot, all, few, ...
Negations	not, no, never, ...

Table 3: LSM categories by LIWC. In some versions there are slight differences regarding the included markers (e.g., in linguistic style coordination Danescu-Niculescu-Mizil et al., 2012, the negation marker is not included).

YAP POS-tags

Tag	Examples of Hebrew Words in Tag (Translation)
ADVERB	כ (about)
AT	את (term used to indicate a direct object)
BN	מתרוצצת (scampering), רוצה (wanting), ...
BNT	לובשי (wearing), ...
CC	כאילו (like), אבל (but), אם (if), ...
CD	אחת (one), 44, ...
CDT	שני (two), ...
CONJ	ו (and)
COP	הייתי (was), היא (is), ...
DEF	ה (the)
DT	איזשהו (some), איזשהי (some)
DTT	שום (any), כל (all), ...
EX	יש (exist), אין (not exist)
IN	בשביל (for), אצל (at), ...
INTJ	נא (please)
JJ	קשה (hard), בטוח (safe), ...
JJT	עומסי (load), ...
MD	נוכל (could), תוכלי (could), צריכה (need), ...
NCD	40, 30%, ...
NEG	לאו (not)
NN	ארץ (country), קניון (mall), משהו (somthing), ...
NNP	חולון (Holon), צרפת (France), ...
NNPT	פלמח (Palmach)
NNT	קרית (a first part in names of cities and neighborhoods), ...
POS	של (of)
PREPOSITION	ל (to), ב (at), ...
PRP	הוא (he), זה (it), אני (I), ...
QW	למה (why), מי (who), איפה (where), ...
RB	רק (only), מאוד (really), מהר (quickly), ...
REL	ש (that)
S_PRN	את (you), היא (she), אני (I), ...
TEMP	כש (when)
TTL	אדון (Mr.), ...
VB	להתלבש (to dress), נפלו (fall), ...
yyCLN	:
yyCM	,
yyDASH	-
yyDOT	.
yyELPS	...
yyEXCL	!
yyLRB	(
yyQM	?
yyQUOT	"
yyRRB	)

Table 4: POS-tags by Hebrew parser YAP.
For the full list and meanings see https://nlp.biu.ac.il/~rtsarfaty/onlp/hebrew/postags