Measuring Linguistic Synchrony in Psychotherapy

Natalie Shapira, Dana Atzil-Slonim, Rivka Tuval Mashiach, Ori Shapira

nd1234@gmail.com
{dana.atzil, tuvalmr}@biu.ac.il
obspp18@gmail.com

Bar-Ilan University, Israel

Abstract

We study the phenomenon of linguistic synchrony between clients and therapists in a psychotherapy process. Linguistic Synchrony (LS) can be viewed as any observed interdependence or association between more than one person's linguistic behavior. Accordingly, we establish LS as a methodological task. We suggest a LS function that applies a linguistic similarity measure based on the Jensen-Shannon distance across the observed part-of-speech tag distributions (*JSDuPos*) of the speakers in different time frames. We perform a study over a unique corpus of 872 transcribed sessions, covering 68 clients and 59 therapists. After establishing the presence of client-therapist LS, we verify its association with therapeutic alliance and treatment outcome (measured using WAI and ORS), and additionally analyse the behavior of *JSDuPos* throughout treatment.

Results indicate that (1) higher linguistic similarity at the session level associates with higher therapeutic alliance as reported by the client and therapist at the end of the session, (2) higher linguistic similarity at the session level associates with higher level of treatment outcome as reported by the client at the beginnings of the next sessions, (3) there is a significant linear increase in linguistic similarity throughout treatment, (4) surprisingly, higher LS associates with lower treatment outcome. Finally, we demonstrate how the LS function can be used to interpret and explore the mechanism for synchrony.[1]

1 Introduction

When people interact, they tend to naturally coordinate their behavior over time. Interpersonal synchrony is defined as the degree to which the behaviors in an interaction are nonrandom and patterned in both timing and form (Bernieri and Rosenthal, 1991). When this pattern occurs, it is often associated with greater rapport between the conversational partners (Butler and Randall, 2013). Research has demonstrated the beneficial effect of synchrony across various interpersonal relationships, such as between spouses or friends, as well as between parents and their children (Feldman, 2012).

The growing acknowledgment of the importance of synchrony in interpersonal relationships has recently led psychotherapy researchers to address the impact of synchrony in the psychotherapeutic process as a way to predict better therapeutic outcomes (Koole and Tschacher, 2016; Paulick et al., 2018).

Recent studies have demonstrated synchrony between clients and therapists through different modalities (Wiltshire et al., 2020). For example, higher levels of body-movement synchrony have been tied to more positive therapeutic relationships and treatment outcomes (Ramseyer and Tschacher, 2011, 2014; Tschacher and Meier, 2020), vocal synchrony was associated with higher empathy ratings (Imel et al., 2014), and physiological arousal coordination has been tied to client-perceived therapist empathy (Marci et al., 2007). However, *linguistic* synchrony (LS) between client and therapist has received relatively little attention.

The words and language clients and therapists use in psychotherapy sessions reflect their internal thoughts and emotions and reveal important information about their interaction. Thus, many of the active ingredients of psychotherapy are found in the words and how they are uttered within psychotherapy sessions. Client and therapist LS may reflect their ability to work together in concert and their adjustment to each other's language over time, which may in turn lead to better therapeutic outcome.

With the increased amount of conversational texts accessible, applying natural language processing is an appealing step for mental health research (e.g. Sharma and De Choudhury, 2018; Zhang

[1] For code availability please contact authors.

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 158 - 176
July 15, 2022 ©2022 Association for Computational Linguistics

and Danescu-Niculescu-Mizil, 2020). Indeed, transcripts of psychotherapy sessions have recently become more readily available thanks to advanced ASR transcription technology. These transcripts allow the analysis of LS in psychotherapy (see Section 2).

The few studies that have considered client-therapist LS have tended to focus on one session, and assessed it's association with therapy processes (e.g., Lord et al., 2015; Pérez-Rosas et al., 2017). The extent to which LS develops from session to session and its association with treatment outcome were yet to be explored in a statistically sound manner. Furthermore, a major criticism on studies on interpersonal synchrony concerns the lack of control for coincidental random synchrony Ramseyer and Tschacher (2010). Based on studies that distinguish genuine synchrony from pseudosynchrony in physiological data, the current study proposes a method to assess LS, that is adapted for sequences of texts (Section 4). Section 5 presents a LS function, inspired by previous work addressing LS.[2] We examine client-therapist LS throughout treatment (N = 74, average number of sessions = 12.56, a total of 872 transcripts), session by session, and the association between LS and treatment process and outcome.

In Section 6 we demonstrate the implications of the ability to measure LS. Synchrony is viewed as an important mechanism of change between the client and the therapist, which leads in turn to a better bond and to a better outcome (for review see Koole and Tschacher, 2016; Paulick et al., 2018). When applying the proposed LS function on our dataset, the method displays an association to quality of client-therapist relationship and treatment outcome (Section 6.1), as well as a significant linear change across treatment (Section 6.2). Additionally, we show how the LS function can be used to interpret and explore the mechanism for synchrony (Section 6.3). Finally, we discuss limitations and potential future work in Section 7.

2 Related Work

We focus on previous work researching LS in psychotherapy.

Lord et al. (2015) dealt with motivational interview training treatment (N=122), where each treatment has a single 20-min transcribed session. They measured synchrony between client and therapist with function word coordination on the ordered utterances in a session (Danescu-Niculescu-Mizil et al., 2012). They show that high empathy sessions display greater coordination of function words compared to low empathy sessions. Overall, average coordination of function words is notably higher in high empathy vs. low empathy sessions.

Pérez-Rosas et al. (2017) explored counseling interaction dynamics (N=276; each session with 5 annotation points) and their relation to counselor empathy during motivational interviewing.

The two latter studies were based on synchrony within a single session. Thus, they could not examine patterns of change across treatment. In addition, while these studies demonstrated the presence of LS in sessions characterized by high empathy between clients and therapists, they do not explore the association between LS and other treatment processes and outcome.

Althoff et al. (2016) measured how various linguistic aspects of written conversations (15,555), as opposed to spoken, correlate with outcomes. This dataset is much larger than in our study, however they analyze the counselor's point of view (N=408) (as opposed to dyads) and overlooked the synchrony across long-term treatment.

Borelli et al. (2019) examine how language style matching (LSM; Niederhoffer and Pennebaker, 2002), clients' relational histories, and symptoms were associated within treatment. On a pilot test using a small sample (N=7, sessions=4) they found that LSM values decrease over the course of treatment, and that greater client interpersonal problems prospectively predict lower early LSM in client-therapist dyads, which in turn predicts greater post treatment psychiatric distress.

Aafjes-van Doorn et al. (2020) demonstrate the clinical usefulness of the LSM and rLSM (Müller-Frommeyer et al., 2019) approach in psychotherapy outcome measures with a small sample (N=7, sessions=20). They also described a case study comparing LSM values to observer-rated measure of working alliance, and conclude that a larger-scale study is required for examining the relationship between synchrony and alliance and outcome.

3 Linguistic Synchrony Definition

Inspired by behavioral and physiological synchrony (Bernieri and Rosenthal, 1991; Palumbo et al.,

[2] As opposed to previous work addressing LS, our LS function does not rely on LIWC (Tausczik and Pennebaker, 2010) since it does not support Hebrew language. See Appendix A.4 for a comparison between the use of LIWC and our method.

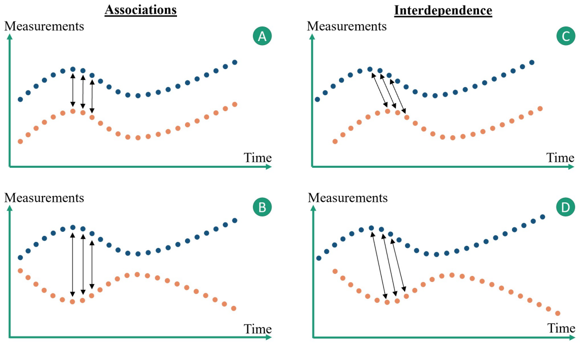

Figure 1: Illustrations of synchrony between repeated measures of two participants (blue and orange) as association (A,B) and interdependence (C,D), with a similarity (A,C) and complementary (B,D) behavior.

2017), **Linguistic Synchrony (LS)** can be viewed as *any observed association or interdependence between people's language dynamics, as indexed by their continuous spoken words, that are nonrandom or patterned in both timing and form.*

Association is a relationship between variables that makes them statistically dependent (e.g., as measured by correlation coefficient see Figure 1.A,B).

Interdependence is the state in which two or more variables rely on or react with one another such that one cannot change without affecting the other (VandenBos, 2007) (e.g., see Figure 1.C,D).

Language dynamics of a conversation are the changes in language use, for each participant individually, that can be captured over time by assessing utterances at a number of time points.

Non-random or patterned associations or interdependences are quantified by adopting an approach by Ramseyer and Tschacher (2010), "surrogate test", that distinguishes genuine synchrony from pseudosynchrony, which may arise due to random coincidence. This definition outlines the statistical tests required to show that a function can indeed measure LS, by pairing texts with non-original replacements and showing a significant difference in the synchrony measure.

4 Formalizing a Task

With respect to the LS definition, we formalize a task, for finding a function that measures LS, as follows:

Given a sample $[[(c^i_j, t^i_j)]_{j=1}^{m_i}]_{i=1}^{n}$ of n pairs (e.g., clients and their therapists) with m_i repeated measures (e.g., sessions in treatment) of lingual texts (i.e., c^i_j, t^i_j are each a written or transcribed text sequence) from a population P, function f : $L \to \mathbb{R}$ is said to be *Measuring Linguistic Synchrony (MLS)* within P, where L is a list of text pairs, if for a set of random texts r^i_j, the sample of values $[f([(c^i_j, t^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$ statistically significantly differs from the generated sample of values $[f([(c^i_j, r^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$ and $[f([(r^i_j, t^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$.[3]

Intuitively, we would like to find a function that is able to recognize that a given list of text pairs has a non-random dependence.

To capture more than just "*any* observed interdependence or association", defined are two additional *pseudosynchrony* tests that use *surrogates* in place of the random texts.

Within challenge: a text c^i_j is paired with a different text contained within t^i's list of repeated measures. Formally, $[f([(c^i_j, t^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$ statistically significantly differs from $[f([(c^i_j, t^i_{l_j})]_{j=1}^{m_i})]_{i=1}^{n}$ where $l_j \in \{1, \dots, m_i\}$ and $l_j \neq j$.

Between challenge: a text c^i_j is paired with any text not contained within t^i's list of repeated measures. Formally, $[f([(c^i_j, t^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$ statistically significantly differs from $[f([(c^i_j, t^{k_j}_{l_j})]_{j=1}^{m_i})]_{i=1}^{n}$ s.t. $k_j \in \{1, \dots, n\}$, $k_j \neq i$ and $l_j \in \{1, \dots, m_{k_j}\}$.

Linguistic synchrony. Populations A and B have different levels of *synchrony* with respect to $f \in MLS$ if f values on population A are statistically significantly different from f values on population B.

Synchrony direction. In order to determine the direction of the synchrony, i.e., whether low or high values of f will be considered as synchrony, we compare the f values of the original sample (i.e., $[f([(c^i_j, t^i_j)]_{j=1}^{m_i})]_{i=1}^{n}$) to the f values of the *surrogates* sample. If f values are lower for the original sample than for the surrogate sample, then lower f values imply higher synchrony. Correspondingly, if the f values are higher in the original sample, then higher f values imply higher synchrony.

Task objective. The objective is to find an MLS function that maximizes the *magnitude* – the strength of synchrony – typically represented by the effect size of the statistical test. In addition, the MLS function should strive to reveal an aspect with which synchrony can be expressed.[4]

[3] In social sciences, as opposed to exact sciences, a measurement is not required to obey a well-defined unit of measure.

[4] An important goal in synchrony research is to provide an interpretation for the observed synchrony. E.g., for synchrony in autism, there are diagnostic tools that assess social skills

We emphasize that synchrony is a change that occurs over time, as opposed to similarity that is measured at a single point. Additionally, synchrony may be expressed through, e.g., complementary behavior (Ackerman and Bargh, 2010; Chartrand and Lakin, 2013) or coordination that can be observed in a non-aligned manner, e.g., shifting content or aggregating several samples together (Figure 1).

Limitation. There exist outlier MLS functions that meet all requirements of the task definition, but do not actually measure synchrony. For example, a function that internally stores the full sample ($[[(c^i_j, t^i_j)]^{m_i}_{j=1}]^{n}_{i=1}$) and simply returns 1 if a given pair ($(c^i_j, t^i_j)]^{m_i}_{j=1}$) appears in the sample and 0 otherwise. A function with a reasonable description length (e.g., memory use) would not allow such functionality. Moreover, proposing such a function does not serve the purpose of synchrony research. Another example is a function that randomly chooses a value that happens to correctly distinguish between an actual pair and a surrogate pair. Such behaviour is not statistically expected.

We next present an LS function that exposes linguistic similarity over time, and in Appendix A.3, a different function that exhibits complementary behavior.

5 Exemplifying Solution

Adhering to the formalized conception of MLS, we next lay out a use case brought from psychotherapy research. First, the data we use is described, then a candidate MLS function is presented, and finally the function is tested for MLS.

5.1 Dataset Description

We employ a dataset of a total of 872 psychotherapy session transcripts, in Hebrew, from 74 different dyads (client-therapist pairs), constructed by 68 clients and 59 therapists. A treatment of a dyad is composed of several sessions (Mean=12.56; SD=4.93). For the purposes of this study, we referred only to verbal text and punctuation, marked by how they were heard (comma as a short pause in speech) and not by how proper sentences should be written.[5] Prior to each session, clients self-reported[6] their functioning, measured using the ORS questionnaire (Miller et al., 2003), which is considered to be an indicator for progress in treatment (see Appendix A.2.1). After each session, therapists and clients reported their perspective for the quality of the relationship during the session, measured by the WAI questionnaire (Horvath and Greenberg, 1989) (see Appendix A.2.2). We note that this dataset is an order of magnitude larger than those used in the few previous works dealing with psychotherapy text analysis (see Section 2).

5.2 Candidate Synchrony Function

Algorithm 1: Lingual distance of client's (c) and therapist's (t) texts list (size=m)

```
1 candidateMLS(c,t,m);
2 for $j \leftarrow 1$ to $m$ do
3     $cPos_j, tPos_j \leftarrow pos(c_j), pos(t_j)$;
4     $cuPos_j \leftarrow prDis(cPos_j)$;
5     $tuPos_j \leftarrow prDis(tPos_j)$;
6     $JSDuPos_j \leftarrow jsd(cuPos_j, tuPos_j)$;
7 end
8 return: $average(JSDuPos)$
```

We present Algorithm 1 as a candidate MLS function.[7] *candidateMLS*, receives as input lists C^d and T^d (d represents specific dyad name) both of size m_d, of a client's and matching therapist's transcribed sessions. The client's and therapist's texts are paired by sessions. I.e., each list element contains the client's or therapist's utterances from a single session, $c^d_j \in C^d$ ($t^d_j \in T^d$) is a concatenation of all client's (therapist's) sentences within session number j, and c^d_j and t^d_j are from the same session, for each session j.

Inspired by previous work addressing LS, the *candidateMLS* function converts each element in the two lists to a probability distribution of unigram part-of-speech (POS) tags (see Appendix A.4 for the relation between LSM categories used in previous works and POS tags). In line 3 of Algorithm 1, the pos function[8] extracts the POS tags from the client's (therapist's) text c_j (t_j) in session j and

such as eye contact or speech turn coordination in conversation. Accordingly, this allows planning respective interventions that address these social skills (e.g., Hopkins et al., 2011). For LS, we provide a possible interpretation in Section 6.3.

[5] For further details about participants, treatment, transcriptions procedure and ethical concerns, see Appendix A.1.

[6] Note that there are biases related to subjective self-reports (Kazdin, 2008). Nevertheless, it is common to build upon such self-reports for psychotherapy research.

[7] An additional candidate synchrony function is presented in Appendix A.3, which measures complementary behavior and applies correlation for computing the magnitude of synchrony.

[8] We used YAP (More and Tsarfaty, 2016) for Hebrew POS tagging.

stores the resulting sequences in $cPos_j$ ($tPos_j$). In lines 4 and 5, the $prDis$ function converts the $cPos_j$ and $tPos_j$ POS sequences to their distributions, and stores them in $cuPos_j$ and $tuPos_j$ respectively. In line 6, the jsd function calculates the Jensen-Shannon Distance[9] (JSD) (Fuglede and Topsoe, 2004) between distributions $cuPos_j$ and $tuPos_j$ (method denoted *JSDuPos*). Finally, *candidateMLS* outputs the average of *JSDuPos*$_j$ values ($j \in [1, m_d]$), providing a synchrony score for dyad d, where a lower score means higher synchrony.

Note the difference between *JSDuPos* and *candidateMLS*. *JSDuPos* is a measure of linguistic **similarity** between the client and the therapist that is calculated for each **session** separately. A lower *JSDuPos* value indicates a closer distance between texts and therefore a higher similarity. *candidateMLS* is a measure of linguistic **synchrony** between the client and the therapist, that is calculated for a **treatment**. A lower *candidateMLS* value indicates lower synchrony (see Section 3 for an explanation on synchrony direction and Section 5.3 on how we determined the direction for our function).

As *JSDuPos* is an interpretable measure of linguistic similarity, it is useful for psychologists to better understand mechanisms of change throughout treatment, i.e., by viewing changes in use of part of speech, as demonstrated in Section 6.3. Furthermore, this function does not require training data, as opposed to data-hungry similarity methods (e.g. Bevendorff et al., 2020; Boenninghoff et al., 2020), which is pertinent in domains where data is rather scarce. Other measures, such as those used for authorship attribution (Koppel et al., 2009; Stamatatos, 2009; Juola, 2008; El and Kassou, 2014), are appealing MLS candidate functions, and we advocate future research to inspect such options.

5.3 Synchrony Function Evaluation

To assess whether the candidate function meets the MLS criteria, we test the *Within* and *Between* challenges, using the corpus of client-therapist conversations from Section 5.1.

The paired sequences of the conversations are as follows: each dyad i ($i \in [1, 74]$) has m_i sessions $S^i_{1:m_i}$. For each session $s^i_j \in S^i_{1:m_i}$ we separated the utterances of the client c^i_j and the utterances of the therapist t^i_j, producing sequences of texts $C^i_{1:m_i}$ and $T^i_{1:m_i}$. The whole corpus can be described as $[[(c^i_j, t^i_j)]^{m_i}_{j=1}]^{74}_{i=1}$.

Within-experiment: (1) For each dyad i: (1.1) Calculate *candidateMLS* on the client's $C^i_{1:m_i}$ and the corresponding therapist's $T^i_{1:m_i}$ to get synchrony magnitude value v^i. (1.2) Choose random permutation $perm(T^i_{1:m_1})$, and calculate *candidateMLS* on the client's $C^i_{1:m_1}$ and $perm(T^i_{1:m_1})$, to get result w^i. Due to non-normally distributed data, (2) Compute Wilcoxon signed-rank one-tail test[10] and Cohen's d (Cohen, 1988) on vectors $V = [v^i]$ and $W = [w^i]$, expecting values of V to be significantly lower than values of W.

This experiment is repeated 100 times on different permutations. All experiments yielded a significant superiority (Dror et al., 2020) of genuine synchrony versus pseudosynchrony (p < 0.05) with a small effect size (average Cohen's d = 0.12). V (M = 0.174; SD = 0.034) exists in a lower level compared to W (surrogate session) (M = 0.179; SD=0.035).

Between-experiment: (1) For each dyad i: (1.1) Compute v^i as described above. (1.2) For each $c^i_j \in C^i_{1:m_i}$, randomly choose, with replacements, a different therapist session t^k_l ($i \neq k$) from the entire set of therapists sessions $[[t^i_j]^{m_i}_{j=1}]^{74}_{i=1}$ and calculate *candidateMLS* on $C^i_{1:m_i}$ with the randomly generated therapist sequence, to get result b^i. (2) Compute Wilcoxon signed-rank one-tail test and Cohen's d on vectors $V = [v^i]$ and $B = [b^i]$.

On 100 different experiments (different replacements), all trials yielded a significant superiority of genuine synchrony versus pseudosynchrony (p < 0.05) with a very large effect size (Sawilowsky, 2009) (average Cohen's d = 1.459). V exists in a lower level compared to B (surrogate therapist) (M = 0.218; SD = 0.028).

Both *Within*- and *Between*-challenge tests pass, indicating that the candidate function meets the MLS criteria. Results are depicted in Figure 2.

[9] Jensen-Shannon Divergence is based on Kullback-Leibler Divergence with a simple manipulation that makes it symmetric (instead of measuring the relative entropy between the two distributions, measure the average of the entropies between each of the distributions and their average distribution) and thus maintains the triangular inequality. JS-Distance is the root of JS-Divergence. We chose *distance* over *divergence* since distance is the more common preference in the literature (1,850,000 search results in Semantic Scholar vs. 239,000).

[10] Based on previous studies, we hypothesize that the synchrony direction is inversely proportional to the similarity in our function. Thus, we expect lower *candidateMLS* values in the original text-pair sample compared to the surrogate sample.

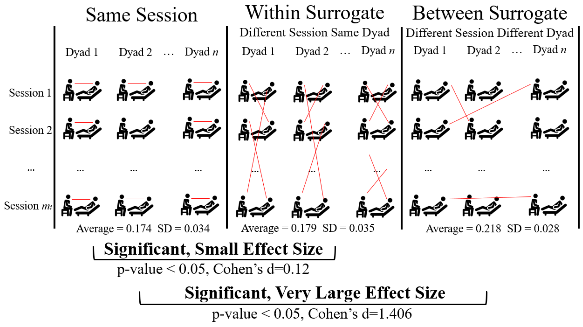
Same Session
Within Surrogate
Different Session Same Dyad
Between Surrogate
Different Session Different Dyad
Dyad 1
Dyad 2
...
Dyad n
Session 1
Session 2
...
Session m_i
Average = 0.174 SD = 0.034
Average = 0.179 SD = 0.035
Average = 0.218 SD = 0.028
Significant, Small Effect Size
p-value < 0.05, Cohen's d=0.12
Significant, Very Large Effect Size
p-value < 0.05, Cohen's d=1.406

Figure 2: The degree of synchrony in conversations between therapist and client compared to pseudosynchrony in conversations that did not take place.

	Session Level			Dyad Level		
Variable	Obs.	M(SD)	Range	Obs.	M(SD)	Range
JSDuPos	871	0.17 (0.05)	0.08-0.9	74	0.17 (0.03)	0.11-0.3
ORS	860	24.4 (7.96)	0.3-40	74	24.5 (6.41)	10.15-38.24
C_WAI	823	50.89 (23.82)	4-84	74	49.48 (23.02)	9.5-84
T_WAI	831	41.69 (18.61)	0-74	74	40.33 (17.88)	8.75-67.61

ORS = Outcome Rating Scale; WAI = Working Alliance Inventory;
C = Client; T = Therapist; Obs. = Observations

Table 1: Descriptive statistics of treatment measurements (processes and outcome) and of our *JSDuPos* function. *JSDuPos* = JS-Distance between probability distributions of unigram POS-tags.

6 Implications of the Candidate Function

Psychology research puts forth much effort in trying to understand the synchrony phenomenon and mechanism (Section 1). In addition, studies show a link between client-therapist synchrony and treatment processes and outcomes (Sections 2). Thus, we examine the relationship between LS and treatment measures through the candidate function (Section 6.1), analyze the change of *JSDuPos* over the course of treatments (Section 6.2), and demonstrate what can be extracted from the function to further understand LS (Section 6.3).

6.1 Associations with Treatment Process and Outcome

Hypothesis 1: We expect that *JSDuPos* and *candidateMLS*, both associate with treatment process and outcome.

(Hypothesis 1a) A lower *JSDuPos* value in a session, i.e., higher linguistic similarity, associates with: (1) a higher level of alliance between therapist and client as reported by both therapist and client at the end of the session, and (2) a higher level of treatment outcome as reported by the client at the beginnings of the current and next sessions. I.e., $JSDuPos(c_s^d, t_s^d)$ correlates with Client_WAI$_s^d$, Therapist_WAI$_s^d$, ORS$_s^d$ and ORS$_{s+1}^d$.

(Hypothesis 1b) A lower *candidateMLS* value of a treatment, i.e., higher LS, associates with: (1) a higher level of alliance between the client and therapist as reported both by client and therapist at the end of each session in the treatment, and (2) a higher level of treatment outcome as reported by the client at the beginning of each session. I.e., $candidateMLS(C^d, T^d)$ correlates with average values of Client_WAI^d, average of Therapist_WAI^d and average of ORS^d.

Results: The descriptive statistics – means, standard deviations and ranges for all the variables – are presented in Table 1.

To examine *(Hypothesis 1a)* we conducted a multilevel model (MLM) test[11] (Bolger and Laurenceau, 2013) that predicts a session's treatment process/outcome value with the corresponding *JSDuPos* (dyad mean-centered) value. Multilevel models allow estimation of two levels (a within-dyad level and a between-dyad level) and accommodate non-balanced data (see Bolger and Laurenceau) as in our case (i.e., sessions nested within dyads and dyads have different numbers of sessions). We used two-level MLM and not three-level MLM (sessions nested within dyads nested within therapists) because of the limited number of clients per therapist.

To examine *(Hypothesis 1b)*, the same multilevel model test factors in the *candidateMLS* value (as a grand mean center of *JSDuPos* dyad values, denoted $meanJSDuPos$).

The mixed-level equation is as follows:

$$\begin{aligned} &Treatment\_Measure_s^d = \\ &(\gamma_0^0 + u_0^d) \\ &+ (\gamma_1^0 + u_1^d) JSDuPos_s^d \\ &+ (\gamma_2^0) meanJSDuPos^d + e_s^d \end{aligned} \tag{1}$$

s.t. *Treatment_Measure* $\in$ {ORS, Client_WAI, Therapist_WAI}. $Treatment\_Measure_s^d$ for a dyad d in session s is predicted by the sample's intercept (γ_0^0), by dyad d's deviation from this intercept (u_0^d), by the average (i.e., fixed) effects (γ_1^0, γ_2^0) of the predictors, by this client's deviation from the fixed effects (i.e., the random effects: (u_0^d, u_1^d)), and by a level-1 residual term quantifying the session's deviation from these effects (i.e., the random effect at level 1, e_s^c).

We note that to examine the prospective association between the MLS and treatment outcome as

[11] Using the R *lme4* library (Bates, 2010), *lmer* function.

reported by the client at the beginning of the *next* session (ORS^d_{s+1}), Equation 1 was computed with the next session index (index $s+1$ instead of s), as follows:

$$\begin{aligned} &ORS^d_{s+1} = \\ &(\gamma^0_0 + u^d_0) \\ &+ (\gamma^0_1 + u^d_1)\mathit{JSDuPos}^d_s \\ &+ (\gamma^0_2)\mathit{meanJSDuPos}^d + e^d_s \end{aligned} \tag{2}$$

As can be seen in Table 2, consistent with *Hypothesis 1a*, a lower *JSDuPos* value (higher linguistic similarity) in a session associates with a higher level of alliance between the client and therapist as reported both by client and therapist at the end of each session (supporting *(Hypothesis 1a)* (1)), and a higher level of treatment outcome as reported by the client at the beginning of the next session (partially supporting *(Hypothesis 1a)* (2)). However, not consistent with *Hypothesis 1b*, a lower *candidateMLS* value (higher linguistic synchrony) in a treatment associates with a lower level of treatment outcome as reported by the client at the beginning of both the current session and the next session. Although the results of the model predicting ORS was statistically significant, the direction was opposite to the hypothesis. In addition, *candidateMLS* did not show associations with alliance of both client and therapist (i.e., *(Hypothesis 1b)* failed to reject the null hypothesis).

6.2 Similarity Increase throughout Treatment

In order to better understand the synchrony mechanism, we examine the change in similarity between client and therapist over the course of a treatment. Since all previous studies that examine LS were based on a single session or a small scale dataset (i.e., could not examine change over time), the following hypothesis will be tested in an exploratory manner.

Hypothesis 2 (exploratory): We expect an increase in linguistic similarity throughout treatment.

Results: To examine the extent in which similarity changes throughout a treatment, a linear growth-curve analysis is conducted over the *JSDuPos* values of treatments.[12] Growth curve models typically refer to statistical methods that allow the estimation of patterns of change over time (the most basic feature of an intensive longitudinal outcome) (Bolger et al., 2003).

Results show a significant linear change across treatment. Specifically, the time trend was negative ($b = -0.001$, $SE = 0.0002$,$t = -4.854$, $p < 0.001$), indicating that on average client-therapist linguistic similarity was higher (*JSDuPos* was lower) in the later stages of therapy compared to the initial stages. See Figure 5 in the appendices for a visualization of the constant decrease in *JSDuPos* over time.

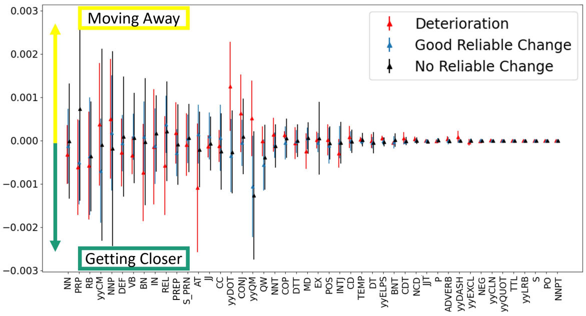

Figure 3: Average and standard deviation of changes in part-of-speech tag frequencies from session to session by all clients and therapists, viewed separately for three groups of dyads divided according to treatment outcome. On average over all treatments with good reliable change, question-mark (yyQM) is the tag for which therapists and corresponding clients move closer the most over a treatment.

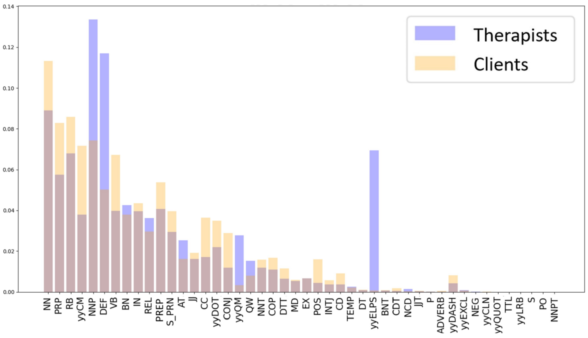

Figure 4: The most asynchronous treatment with frequencies of POS tags in a client's (orange) and therapist's (purple) transcriptions. The three major sources of asynchrony, with the highest frequency gap, are the parts-of-speech NNP (proper noun singular), DEF (morphological determiner) and yyELPS (ellipsis).

6.3 Utility of LS in Treatment

In this section we will demonstrate how the LS mechanism can be further explored.[13] As shown in Section 6.2, *JSDuPos* values decrease over treatment. We explore in what sense the client and therapist become closer in terms of changes in POS

[12] Using the R *nlme* library, *lme* function.

[13] There are no clinical recommendations here but rather a demonstration of the benefits of an interpretable synchrony function. In the current study it is not possible to examine the causality relation between synchrony and outcome.

	Previous Week ORS		Next Week ORS		Client_WAI		Therapist_WAI	
Predictors	Estimates (Std. Err)	95% CI (t value)	Estimates (Std. Err)	95% CI (t value)	Estimates (Std. Err)	95% CI (t value)	Estimates (Std. Err)	95% CI (t value)
(Intercept)	24.50*** (0.711)	[23.11, 25.90] (34.439)	24.71*** (0.744)	[23.26, 26.17] 33.22	49.50*** (2.674)	[44.26, 54.74] (18.51)	40.36*** (2.077)	[36.29, 44.43] (19.433)
Session *JSDuPos*	-8.41 (70171)	[-22.46, 5.65] (-1.172)	-17.16*** (4.868)	[-26.70, -7.62] (-3.525)	-30.90*** (8.811)	[-48.17, -13.63] (-3.507)	-25.76*** (7.446)	[-40.35, -11.17] (-3.46)
Dyad *meanJSDuPos*	64.54*** (21.987)	[21.44, 107.63] (2.935)	47.74* (21.350)	[5.89, 89.58] (2.236)	-72.70 (83.612)	[-236.58, 91.17] (-0.87)	-79.80 (65.122)	[-207.44, 47.84] (-1.225)
Observations Conditional R (ICC)	859 0.646 (0.62)		849 0.680 (0.67)		822 0.931 (0.93)		830 0.920 (0.92)	

Note. ***p<0.001; **p<0.01; *p<0.05; ORS = Outcome Rating Scale; WAI = Working Alliance Inventory
JSDuPos = Jensen–Shannon-Distance between Probability Distribution over Unigram POS-tag;
meanJSDuPos = Result of the synchrony function (*candidateMLS*) which is the average *JSDuPos* for each dyad.

Table 2: Associations between similarity (*JSDuPos*) or synchrony (*meanJSDuPos*), and treatment measurements – outcome (ORS) or process (WAI).

tag distributions over a treatment, using two approaches.

In the first approach, we analyze the changes in use of POS tags in treatment in three different groups of dyads (of the 74 available): those with a good reliable change in treatment, those with a reliable deterioration, and those with no reliable change.[14] Then, for each POS tag p and for each dyad d in its group, for a sequence of sessions $s_1^d, s_2^d, ..., s_{n_d}^d$ we compute the distances $\delta_1^{d,p}, ..., \delta_{n_d}^{d,p}$ where $\delta_i^{d,p}$ is computed as the absolute difference between the client's frequency of p and the therapist's frequency of p in session i. We then compute the difference in distances between consecutive sessions of the treatment $\Delta_i^{d,p} = \delta_i^{d,p} - \delta_{i-1}^{d,p}$. The score for this treatment and POS tag is then $score^{d,p} = \sum_2^{n_d} \frac{\Delta_i^{d,p}}{n_d - 1}$, i.e., the average of the differences in the sequence of sessions. Finally, for each POS tag separately, we calculate the average and standard deviation of scores of all dyads within their group. A lower value for a POS tag means the clients' and corresponding therapists' tag frequency becomes more similar overall.

As seen in Figure 3, in the dyads with a good reliable change, the POS tag frequencies of clients and therapists moved towards each other in the question-mark (yyQM) and question (QW) tags. When zooming in from part-of-speech- to the lexical-level, i.e., analysing frequencies of question *words*, we found the biggest change in the "what" token. Throughout the treatment, the frequency of "what" increases for clients ($+0.1\%$) while decreasing for therapists (-0.1%). See also Figure 6 in the Appendices for separate client and therapist points of view of a similar analysis.

Another approach for exploring the LS mechanism is by analyzing the contributors that influence the magnitude of synchrony within a specific treatment. We demonstrate this through a case study from our data in the treatment with the lowest synchrony value as calculated with *candidateMLS* (highest average *JSDuPos* scores). This treatment was also considered unsuccessful as measured by ORS. Figure 4 shows the POS tag distribution of the whole treatment for a client-therapist dyad. The differences in the tag distributions may hint at reasons for the unsuccessful treatment. Here we see that the therapist uses some POS tags far more often than the client. For example, there is a frequent use of ellipses (yyELPS), indicating many silent moments. Accordingly, these tags can expose behavior that may have gone unnoticed.

7 Discussion and Future Work

In Section 5.2 we propose a function that is able to measure LS, based on a similarity approach. Future work may assess LS functions that apply different similarity methods. Additionally, new LS functions should examine other forms of synchrony such as coordination and accommodation.

The field of *Authorship Attribution* (Koppel et al., 2009; Stamatatos, 2009; Juola, 2008; El and Kassou, 2014), for example, may inspire development of new LS functions. This field relies on features of complexity measures (e.g., average word length, average number of words in a sentence),

[14] The ORS has a Reliable Change Index (RCI = 5 points) that identifies when change is clinically significant and attributable to therapy. (Low et al., 2012)

syntax, taxonomies, morphological analysis, orthographic/syntactic errors, idiosyncrasies and others. These may be adapted for measuring LS as well.

We note that in this work we describe synchrony as it is commonly referred to in psychology. This definition does not discriminate between *intrinsic* synchrony and *extrinsic* synchrony. Two bodies synchronize intrinsically when they directly influence each other. For example, the moon's motion synchronizes with sea levels due to the gravitational force exerted by the moon on the sea. In other cases, an *external* constituent impacts the two bodies in such a way that they synchronize independently. For example, two clocks are in synchrony with each other as a result of the time specified by an independent source. In the case of LS, the use of linguistic features by two "synchronized" speakers may be due to an outside cause, like a seasonal use of words. When discovering synchrony with a measuring function, the underlying root cause remains unknown. More research should be conducted in order to reveal the confounding variables of synchrony.

8 Conclusion

Researching synchrony enhances our understanding of the mechanisms of change in psychotherapy treatment. Language, in particular, reveals important information about the interaction between a client and a therapist. Following previous work on synchrony research, we formally define a task for measuring *linguistic* synchrony, and describe two tests for quantifying the quality of a function that measures LS. We suggest a function, consisting of a similarity component inspired by methods used in Psychology research, that satisfies the definition and tests. The function and its component are shown to correlate with measures of psychotherapy process and outcome and show a significant linear increase across treatment. Furthermore, we demonstrate how this function can be interpreted for understanding the interaction between the client and therapist throughout treatment. While this nonstandard task of Linguistic Synchrony can strongly contribute to analysis in Psychology, we also generally see it as an intriguing challenge to undertake in comparative textual analysis.

9 Ethical Considerations

This study was approved by an Institutional Review Board and was conducted ethically in accordance with the World Medical Association Declaration of Helsinki. The procedures were part of the routine assessment and monitoring process in the clinic. Informed written consent was obtained from all participants at the outset of this study. Participants are asked to provide written consent that their data will be used for research. They are informed that at any time they may request to terminate their participation in the research and / or request that the content of the recordings be deleted without jeopardizing treatment. All data collected was anonymized and only then exposed to a very small number of researchers, as agreed upon by the participants. More information is avaialbale in Appendix A.1.

Acknowledgements

We would like to thank Yoav Goldberg, Daniel Juravski, Tslil Ofir, Ayelet Meiri, Adar Paz, Dana Stolowicz-Melman, Yarden Menashri Sinai, Almog Simchon, Rotem Dror, Victoria Basmov, Shoval Sadde, Reut Tsarfaty, Nir Milstein, Oded Mayo, Alon Tomashin, Maya Sabag, Shahar Siegman, the synchrony PhD forum members in the Psychology department, and BIU-NLP lab members for helpful discussions and contributions, each in their own way. We thank the anonymous reviewers for their insightful comments and suggestions. This project was partially funded by the Israel Science Foundation (grants 1348/15 and 1278/16); the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 802774 (iEXTRACT); and the Computer Science department of Bar-Ilan University.

References

Katie Aafjes-van Doorn, John Porcerelli, and Lena Christine Müller-Frommeyer. 2020. Language style matching in psychotherapy: An implicit aspect of alliance. *Journal of Counseling Psychology*, 67(4):509.

Joshua M Ackerman and John A Bargh. 2010. Two to tango: Automatic social coordination and the role of felt effort.

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.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Douglas M Bates. 2010. lme4: Mixed-effects modeling with r.

Frank J Bernieri and Robert Rosenthal. 1991. Interpersonal coordination: Behavior matching and interactional synchrony.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. 2020. Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–383. Springer.

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.

Benedikt Boenninghoff, Julian Rupp, Robert M Nickel, and Dorothea Kolossa. 2020. Deep bayes factor scoring for authorship verification. *arXiv preprint arXiv:2008.10105*.

Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003. Diary methods: Capturing life as it is lived. *Annual review of psychology*, 54(1):579–616.

Niall Bolger and Jean-Philippe Laurenceau. 2013. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Jessica L Borelli, Lucas Sohn, BingHuang A Wang, Kajung Hong, Cindy DeCoste, and Nancy E Suchman. 2019. Therapist–client language matching: Initial promise as a measure of therapist–client relationship quality. *Psychoanalytic Psychology*, 36(1):9.

Emily A Butler and Ashley K Randall. 2013. Emotional coregulation in close relationships. *Emotion Review*, 5(2):202–210.

Tanya L Chartrand and Jessica L Lakin. 2013. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.

J Cohen. 1988. Statistical power analysis for the behavioral sciences, 2nd edn. á/l.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2):1–116.

Sara El Manar El and Ismail Kassou. 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12).

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.

Ruth Feldman. 2012. Bio-behavioral Synchrony: A Model for Integrating Biological and Microsocial Behavioral Processes in the Study of Parenting. *Parenting*, 12(2-3):154–164.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Ingrid Maria Hopkins, Michael W Gower, Trista A Perez, Dana S Smith, Franklin R Amthor, F Casey Wimsatt, and Fred J Biasini. 2011. Avatar assistant: improving social skills in students with an asd through a computer-based intervention. *Journal of autism and developmental disorders*, 41(11):1543–1555.

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

Zac E Imel, Jacqueline S Barco, Halley J Brown, Brian R Baucom, John S Baer, John C Kircher, and David C Atkins. 2014. The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of counseling psychology*, 61(1):146.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Patrick Juola. 2008. *Authorship attribution*, volume 3. Now Publishers Inc.

Alan E Kazdin. 2008. Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American psychologist*, 63(3):146.

Sander L Koole and Wolfgang Tschacher. 2016. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology*, 7:862.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

DC Low, SD Miller, and B Squire. 2012. The outcome rating scales (ors) & session rating scales (srs): Feedback informed treatment in child and adolescent mental health services (camhs). *Norwich: Norfolk & Suffolk NHS Foundation Trust.*

Carl D Marci, Jacob Ham, Erin Moran, and Scott P Orr. 2007. Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of nervous and mental disease*, 195(2):103–111.

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, page 273.

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.

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

Lena C Müller-Frommeyer, Niels AM Frommeyer, and Simone Kauffeld. 2019. Introducing rlsm: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, 51(3):1343–1359.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. 2017. Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review*, 21(2):99–141.

Jane Paulick, Anne-Katharina Deisenhofer, Fabian Ramseyer, Wolfgang Tschacher, Kaitlyn Boyle, Julian Rubel, and Wolfgang Lutz. 2018. Nonverbal synchrony: A new approach to better understand psychotherapeutic processes and drop-out. *Journal of Psychotherapy Integration*, 28(3):367.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086.*

Fabian Ramseyer and Wolfgang Tschacher. 2010. Nonverbal synchrony or random coincidence? how to tell the difference. In *Development of multimodal interfaces: Active listening and synchrony*, pages 182–196. Springer.

Fabian Ramseyer and Wolfgang Tschacher. 2011. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79(3):284.

Fabian Ramseyer and Wolfgang Tschacher. 2014. Nonverbal synchrony of head-and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in psychology*, 5:979.

Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26.

Natalie Shapira, Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Dana Stolowicz-Melman, Adar Paz, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, et al. 2021. Hebrew psychological lexicons. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 55–69.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

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*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

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

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.

Wolfgang Tschacher and Deborah Meier. 2020. Physiological synchrony in psychotherapy sessions. *Psychotherapy Research*, 30(5):558–573.

Gary R VandenBos. 2007. *APA dictionary of psychology.* American Psychological Association.

Travis J Wiltshire, Johanne Stege Philipsen, Sarah Bro Trasmundi, Thomas Wiben Jensen, and Sune Vork Steffensen. 2020. Interpersonal coordination dynamics in psychotherapy: A systematic review.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. *arXiv preprint arXiv:2005.04245*.

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/

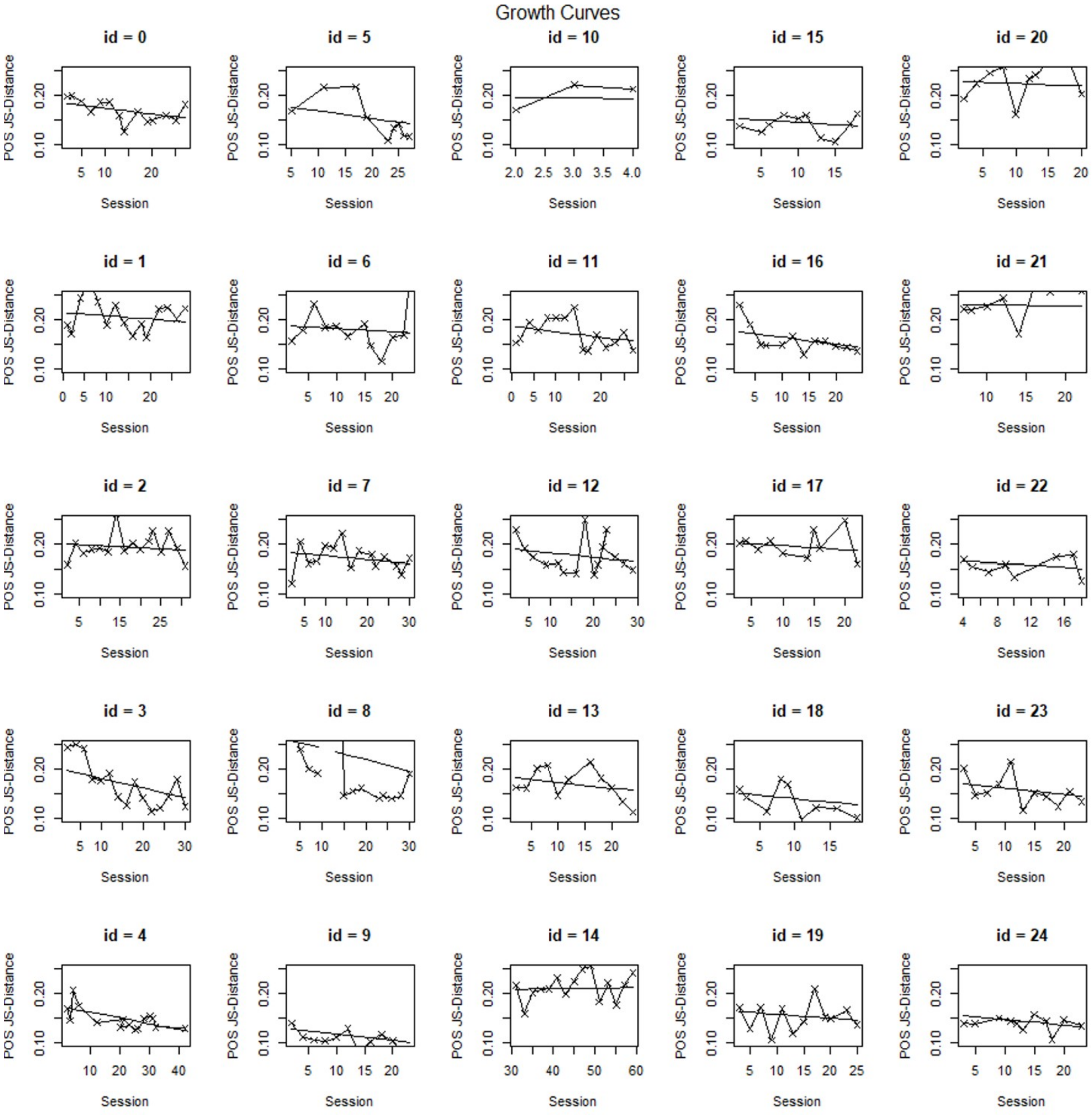

Figure 5: Growth Curves of 25 sampled dyads of the 74 available. There is a decrease of 0.001 units (i.e., slope) of JS-Distance between Probability Distribution over Unigram POS-tag in each session throughout treatment, indicating an increase in linguistic similarity. Results are statistically significant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

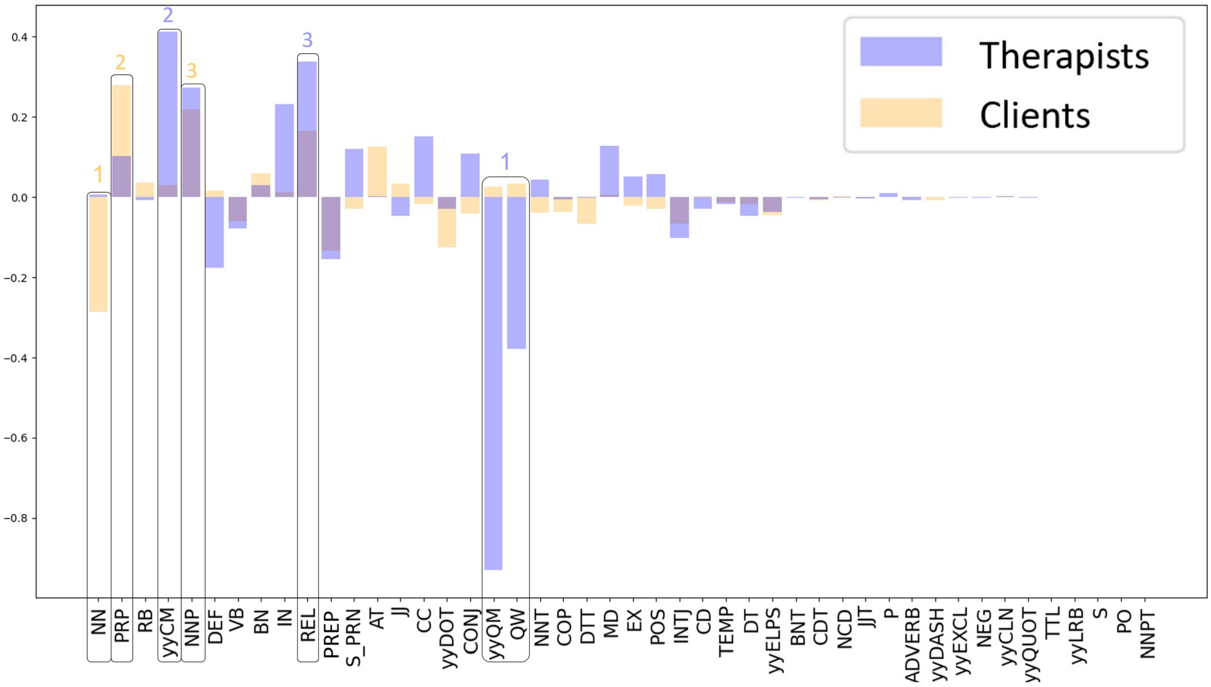

Figure 6: The sum of POS tag frequency changes between consecutive sessions for all clients (orange) and therapists (purple). A positive (negative) value means an overall increase (decrease) in the frequency of a POS tag throughout treatments. The three major changes in treatments for therapists are (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