Preparing Data from Psychotherapy for Natural Language Processing

Margot Mieskes, Andreas Stiegelmayr
Hochschule Darmstadt
Darmstadt, Germany
{firstname.lastname}@h-da.de

Abstract

Mental health and well-being are growing issues in western civilizations. But at the same time, psychotherapy and further education in psychotherapy is a highly demanding occupation, resulting in a severe gap in patient-centered care. The question which arises from recent developments in natural language processing (NLP) and speech recognition is, how these technologies could be employed to support the therapists in their work and allow for a better treatment of patients. Most research in NLP focuses on analysing the language of patients with various psychological conditions, but only few examples exist that analyse the therapists behavior and the interaction between therapist and patient. We present ongoing work in collecting, preparing and analysing data from psychotherapy sessions together with expert annotations on various qualitative dimensions of these sessions, such as feedback and cooperation. Our aim is to use this data in a classification task, which gives insight into what qualifies for good feedback or cooperation in therapy sessions and employ this information to support psychotherapists in improving the quality of the care they offer.

Keywords: psychotherapy, speech resources, German, data preparation, guidelines, data preprocessing

1. Introduction

Mental health care is a very demanding occupation, where quality assurance and improvement is very time- and resource intensive. Improving mental health care through psychotherapy involves two aspects: First, the point of view of the therapist, who is required to constantly attend further education measures, to ensure the quality of their work remains high and by improving it through supervision. Second, the patients' point of view, who has a right for high quality care, especially for conditions, where the treatment greatly benefits from psychotherapy. But due to a lack of therapists, these patients are still treated with drugs, which are less effective. As there is a lack in further education options for therapists, the number of qualified therapists is hardly increasing, resulting in a vicious circle. This is especially apparent in the treatment of patients with schizophrenic conditions. To improve the situation, psychological therapists are collecting data from therapy sessions, which are then judged by experts from their field. Their guidelines define 14 quality dimentions, such as *Feedback* or *Cooperation*, rated on a seven-point Likert scale. This work is currently carried out manually and is very time-consuming and resource extensive. Previous work on applying Natural Language Processing (NLP) shows that NLP can be used to extract certain aspects in human-human communication. In our work, we build on this previous work to support the analysis of the therapy session. Before this, the data has to be processed, which requires methods and tools, we partially develop ourselves. Our work describes efforts in creating a pipeline for preprocessing therapy recordings to use NLP and machine learning to support psychotherapists in analysing their work. This preprocessing itself poses several challenges: As patients are quite sensitive to their surroundings, the therapy sessions should not be disturbed by technical equipment. The solution found by psychologists is using mobile phones as recording devices, which results in a low recording quality. Second, the material is very sensitive and has to be processed accordingly. And finally, most NLP tasks work on very fine-grained parts of language and speech, wheras the judgement of the therapy sessions by experts is on a very high, session-wide level, rather than a single part. In order to be able to apply NLP to this type of data, we present our guidelines for the transcription, which take into account the high level judgements, as well as pointers from the literature on indicators for (for example) good feedback or cooperation. Additionally, we describe our pipeline to extract a range of features for the classification. Previous work in this area has been primarily carried out on English and there is very little information on how applicable the results are to other languages such as German, which we study here. Even though we cannot publish the original recordings due to data protection laws, we will release our guidelines and implementations for further usage.[1]

2. Related Work

As our project is heavily motivated by research results from psychology research, we present not only previous work on NLP in this context, but also results from research on treating psychological conditions, with a focus on the treatment of schizophrenic patients.

2.1. Research in Psychology

Patients with schizophrenia, who suffer from persecution complex and hearing voices are normally treated in therapeutic institutions, using antipsychotic drugs. Only about 1% of the patients are treated with psychotherapy on top (Görgen and Engler, 2005). But there are numerous studies which indicate that psychotherapy is beneficial for this group of patients (see for example (Lincoln, 2014)). At the same time other studies show that therapists lack possibilities for advanced training in this specific area (Lincoln et al., 2014; Mehl and Lincoln, 2014), which leads to a vicious circle, as there are too few therapists for this patient group.

[1] `https://github.com/mieskes/Paranoia/`

This circle is hard to break and this reduces the chances of patients receiving the best possible care.

2.2. Research in NLP

In recent years, NLP has been applied to the context of psychotherapy, specifically focusing on spoken language. There are efforts in finding acoustic correlates for specific emotions in the spoken language (see for example (Schuller and Batliner, 2013)). Additionally, work on finding correlates for stress in general (Paul et al., 2015) or cognitive stress (Hecht et al., 2015) has been carried out. In most cases both acoustic features, as well as linguistic features are used. These elements correlate to quality dimensions as annotated by psychotherapists. But so far, except for general work on emotion detection, work has been primarily carried out on English data. Work on German is rare.

Flekova et al. (2015) work in the educational domain and look at the topic of the quality in classroom interaction. This bears some similarity to determining the quality in therapy sessions and also the quality dimensions relevant for educational researchers and psychologists share common notions, such as *Feedback* and *Cooperation*. Their findings indicate that how participants phrase their utterances and what kind of words they use are indicative of the quality of the interaction.

Chakravarthula et al. (2015) look at the behavior of the therapist, rather than the patients, especially their empathy level in the context of addiction counseling. The authors model the behavior and find that they are able to reliably classify the empathic state based on the output of an automatic speech recognition system.

2.3. NLP and Schizophrenia

Recently, there has been some work on analyzing the language in the context of schizophrenia. Howes et al. (2013) look at the distribution of topics in therapy sessions of schizophrenic patients and relate them to the therapy outcome. They use Latent Dirichlet Allocation (LDA) to determine the topics with respect to hand-coded topics.

Mitchell et al. (2015) analyse data from social media based on information given by the users themselves. Therefore, the authors mention that the data and the results have to be treated cautiously. Nevertheless, they find that various features in the language differed from the language of people in a control group. The authors find differences in the character n-grams, the distribution of topics, measured using Latent Dirichlet Allocation (LDA) and also the usage of LIWC categories differ between the schizophrenic users and the control group.

Kayi et al. (2017) also use social media data in addition to essays written both by patients and control persons on two topics: *their average Sunday* and *what makes them angriest*. The authors look at a range of syntactic, semantic and pragmatic features in the writings of all persons. The authors report that among the syntactic features, Part-of-Speech tags are very predictive, but they also reveal some problems, as the tag for foreign words (FW) primarily marked misspelled words. Among the semantic features, clusters based on word vectors, topics based on LDA and semantic role labelling (SRL) achieve good results. Among the pragmatic features, sentiment intensitiy features perform better than sentiment features.

3. Background on the data collection

We are collaborating with a psychotherapeutic study, where currently data is being collected from real therapeutic sessions. The goal of that project is to prove the effectiveness of a specific therapy method. Therefore, patients are aware that they are being recorded and that the material is used for research. The patients, as well as the therapists explicitely gave their consent to being recorded and the data being used for research purposes. The sessions are then analyzed and classified by at least two trained psychotherapists using a rating manual which describes 14 quality criteria on a seven-point Likert scale (Lecomte et al., 2017). In our initial study, we focus on few of these criteria, which are also reflected in the language used, such as *Feedback*, *Positive Focus* and *Collaboration*, which we briefly describe in the following. As pointed out by Mitchell et al. (2015) with respect to their data, it is possible that the data of this study is also not representative. Both patients and therapists participate voluntarily. This might skew the distribution towards very motivated patients and therapists. Additionally, the recording might have a slight influence on their behaviour, which the data collectors tried to keep to a minimum.

We received the recordings of the sessons via encrypted, password-protected hard drives to reduce the risk of the data being compromised. The recording quality is very poor, as the recording device was a mobile phone, placed on the table, which is occasionally moved, papers are put on top, etc. which disturbs the recordings. Additionally, the data was stored in an .mp3-format, which reduces the quality even further.[2] We did not receive any meta data about the persons in the recordings, therefore, we did not know any personal details about them. Nevertheless, as patients give details about their situation, the data is too sensitive to be made publicly available.

3.1. Quality Dimensions

In our work we use a subset of the 14 quality dimensions. These seven dimensions are described in the following. More details on these and the remaining dimensions can be found in the original manual (Lecomte et al., 2017).

Positive Focus is based on the observation that patients often face prejudices, which focus on their deficiencies or handicaps. The therapist is required to focus on the positive aspects of the person, their strengths and goals, rather than their problems. Facts can be used in order to inform the patient about their symptoms. The seven-point Likert scale used for the rating, states that 0 means that the therapist only focuses on the problems and symptoms and completely neglects goals or strengths of the patient, whereas 6 means that the therapist is aware of the strengths and goals of the patient and supports him/her in finding new strengths and further develop existing ones.

[2] Improving the recording quality by using dedicated recording devices and head-mounted microphones would disturb the therapy session too much. (Personal communication with one of the lead researcher in the data collection study.)

Feedback goes back and forth between patient and therapist. The latter has to ensure, that the patient is involved in the therapy. But the patient should not feel evaluated or judged in a school-like way, but rather that feedback is supposed to be an exchange of points of view in order that both participants understand each other. A rating of 0 states that there is no feedback, neither does the participant ask for it, nor does the therapist ask for the patients approval. If feedback is requested, it is ignored by the therapist. A rating of 6 means that feedback is given regularly and effectively. If the patient reacts unexpectetly the therapist in turn reacts in a positive way and does not discard it as a misunderstanding. Feedback feels natural.

Cooperation is required to successfully treat a patient. It often occurs that patients withdraw and behave passively. The therapist is required to encourage the patient to cooperate and become more active. A rating of 0 indicates that the therapist does not involve the patient in any decision making process and does not motivate him/her to cooperate. A rating of 6 indicates that both parties participate in the session, if there are any problems the therapist reacts sensibly and both come to a decision.

Access to Emotions which means that people can regulate their emotions, which is very difficult for patients with schizophrenia. It is important that the therapist is able to name emotions that show themselves in the here and now. A rating of 0 indicates that the therapist does not even try to name emotions. A rating of 6 indicates that the therapist looks at emotions from different perspectives not only in the here and now, but also in the past.

Identification of Cognition helps the patient to develop a link between their thoughts and their emotions and behaviour in order to change it. The therapist should understand which thoughts are central. A rating of 0 indicates that the therapist does not try to discuss any thoughts or point them out. A rating of 6 indicates that the therapist deals with current lines of thought and supports the patient in identifying behvioural elements in specific situations.

Agenda checks whether there is one or not and whether a previously set agenda actually matches the patients needs. As patients often face difficulties remembering, the agenda should also track the past. A rating of 0 means that there is no agenda or it was not mentioned. A rating of 6 means that an agenda was mentioned and contains all relevant aspects. The patient and the therapists agreed on the set agenda.

Identification of Behaviour is important as the therapist should support the patient in seeing connections between thoughts and behaviour. The therapist should focus on behavioural aspects that hinder success. A rating of 0 indicates that the therapist does not even try to discuss the patients' behaviour. A rating of 6 means that the therapist frequently considers current behaviour and tries to support the patient in connecting emotions and thoughts with his/her current behaviour.

3.2. Manual Annotation

Table 1 shows the distribution of the quality criteria analysed. In total 35 sessions were analysed, although 4 are missing for *Identification Cognition* and *Access to Emotions*. The distribution is skewed towards the higher ratings. No session was rated with 0 for any of these categories and only few have been marked as 1 or 2. The annotators gave a rating of 4 and more for most of the sessions on most of the quality criteria. An exception to this is *Identification Cognition* where 14 sessions were marked with 3.

Quality	1	2	3	4	5	6
Agenda	1	1	0	10	15	8
Positive Focus	0	1	5	9	11	6
Feedback	0	1	4	7	14	9
Cooperation	0	0	2	8	12	13
Access to Emotions	0	3	7	8	8	5
Identification Cognition	0	2	14	7	6	3
Identification Behaviour	0	2	8	9	11	1

Table 1: Frequency of ratings.

We analyse the annotation using DKPro Agreement (Meyer et al., 2014). The agreement on the categories we consider varies greatly. While *Feedback* achieves a fairly high Fleiss' κ score ($\kappa = 0.497$), the categories *Identification of Cognition* and *Identification of Behaviour* achieve a κ of close to 0. One of the reasons is that not all sessions were annotated by at least 2 annotators. Looking in detail at the annotation, we observe that most differences occur in the range of one point (i.e. 3 vs. 4), which are probably hard to distinguish in a task that is fairly hard to begin with.

4. Processing

The first step in processing the data is the transcription. We experimented with off-the-shelf automatic speech recording tools, but the recording quality proved to be too low. Therefore, we have to first transcribe the data manually, for which a tool and guidelines are required.

4.1. Evaluation of Transcription Tools

We put together a list of features the tools had to have. This includes information on the operating system it ran, the quality of the documentation, the graphical user interface, but also import and export file formats. After a preselection process, we examine three tools according to our requirements (ELAN[3] (Wittenburg et al., 2006), OpenSmile[4] and Audacity[5]). We discover a lot of similarities between the tools. Major differences are in the area of supported file formats and the views offered by the tools. Some tools offer unique characteristics, such as the possibility to segment the data. As the objective analysis based on the features did not allow for a clear distinction, we did a preliminary study with a sample of the data and found, that ELAN offered the best usability and therefore, decided to use it.

4.2. Transcription Guidelines

Based on previous work in quality estimation and information in the quality annotation guidelines used in analysing

[3] `https://tla.mpi.nl/tools/tla-tools/elan/`

[4] `http://audeering.com/technology/opensmile/`

[5] `http://www.audacity.de/`

the therapy sessions, we develop transcription guidelines to ensure that the transcription captures phenomena which would be helpful for the classification of quality dimensions in psychotherapy. Additionally, we address specific phenomena, which we observe in our data, such as the usage of dialectal elements, which have to be treated accordingly.[6]

Dialect We translate dialectal elements to standard German, as most NLP tools can hardly deal with dialectal speech. During transcription, we still keep the verbatim expressions, but add a translation to standard German to normalize for varying dialects.

Non-verbal elements We transcribe non-verbal elements in the recordings. This includes elements from social noise, such as variants of *hm*, which can express acknowledgement, question, surprise etc. We also consider elements such as crying and laughter. To analyse the duration of the noise we are marking them via time stamps.

Pauses Pauses are marked also with respect to their length to differentiate between short and long pauses.

Disfluencies Hesitations and disfluencies indicate various emotional states (nervousness, insecurity, etc.). Therefore, we transcribe hesitations, repairs etc. with great detail, following earlier work by Heeman and Allen (1999)

Citations If the patient or therapists cites him-/herself or somebody else, this is transcribed and marked accordingly, to not confuse the words used by a third person with the person saying them.

Punctuation As spoken language not always uses sentences in a traditional sense, we refrain from using a full stop to indicate the end of a sentence. Rather, commas indicate clauses. We use full stops only for obvious sentence ends. Question marks indicate a question, also for rethorical questions. Exclamations marks indicate stressed words. A colon indicates elongated words.

Not understandable elements Due to the recording quality, we have to take into account that elements of the recordings might not be understandable. The annotators mark those elements accordingly. If they can understand something, they can put the most likely transcription there, but nevertheless, keep the non-understandable marking. These elements can then be re-checked and subsequently used cautiously.

4.3. Manual Preprocessing

As of today, we have transcribed over two hours of recordings, which are almost equally split between the therapists and the patients (see Table 2). We see that while therapists have more segments than patients, the patient segments are considerably longer, indicating that they contribute a lot to the conversation. Pauses are on average fairly short, considering the amount of over 400 pauses distributed over patients and therapists. Patients use slightly more pauses than therapists, which are shorter than therapists pauses, suggesting that therapists are careful about their wording. Dialectal elements are more frequently used by patients rather than therapists, indicating that they do not restrict themselves in the way they express themselves.

Observation	Amount
# Segments (Patient)	720
# Segments (Therapist	948
avrg. length of segments (Patient)	00:00:05.355
avrg. length of segments (Therapist)	00:00:03.866
# of disfluencies (Patient)	85
# of disfluencies (Therapist)	56
# of pauses (Patient)	270
# of pauses (Therapist)	146
avrg. length of pauses (Patient)	00:00:01.837
avrg. length of pauses (Therapist)	00:00:02.123
# dialectal elements (Patient)	911
# dialectal elements (Therapist)	535

Table 2: Statistical Information on the current data set.

We observe that the patients have considerably more segments that contain not understandable elements (approx. 30%), wheras the therapist only has about 18% not understandable elements. This supports our approach of focusing on the therapist for the analysis of the therapy quality.

5. Classification

In the following, we describe our pipeline, the features we extract and some preliminary classification results.[7] Our primary tools are DKPro Core[8] for the linguistic preprocessing and Weka[9] for the machine learning.

5.1. Feature Extraction

In order to do meaningful classifications using machine learning methods, we need features extracted from the data. These features are based on the literature presented in Section 2. above and on the guidelines by the psychotherapists (see Section 3.). Therefore, we focus on elements such as pauses, social noise, disfluencies, but also on the vocabulary used. Earlier work indicates, that for example words from specific word groups indicating insight, understanding, etc. point to a good cooperation and a well established feedback culture (Flekova et al., 2015). Among the features we extract, are surface features, which represent the content of each speakers' part in the conversation. This includes number of sentences, but also pauses and their lengths. As Kayi et al. (2017) shows that Part-of-Speech (POS) tags are very predictive to determine the patients, we use syntactic features such as POS, but also the number of questions and the usage of tense (future, past, present). We also look in detail at hesitations and stuttering.

5.2. Preliminary Results

In an initial experiment we distinguish therapist from patient. This is motivated by the final classification, which looks at the interaction between the two, but with a focus on the therapist to find correlates to therapy quality until the recording problems are solved. A first experiment using the full feature set resulted in an accuracy of 73.6.%
Table 4 presents results on our initial experiments using tenfold cross-validation. Surface features, such as segments

[6] The guidelines can also be found in our Github repository.

[7] The implementation is also available from our github repository.

[8] https://dkpro.github.io/dkpro-core/

[9] http://www.cs.waikato.ac.nz/ml/weka/

Feature Set	Precision	Recall	F-Measure
Full	0.742	0.737	0.735
surface	0.732	0.725	0.723
disfluencies	0.492	0.493	0.474
syntactic	0.624	0.624	0.624
segment length only	0.695	0.678	0.671

Table 3: Results of a preliminary classification experiment distinguishing between patients and therapists.

Feature Set	Precision	Recall	F-Measure
Full	0.774	0.775	0.774
Top 10	0.783	0.782	0.782
Stutter	0.598	0.598	0.596
Speech Break	0.590	0.590	0.590
Noise	0.588	0.588	0.584
Social noise	0.618	0.614	0.615
LIWC Data	0.726	0.711	0.694
Dialect	0.566	0.567	0.566
Sentiment	0.581	0.583	0.582

Table 4: Results of Random Forest Algorithm with 10-fold cross validation.

and token lengths allow for a good distinction between patient and therapist – even slightly better than for the full feature set. Interestingly, the features based on disfluencies give the worst results, which are well below chance. Syntactic features give results better than chance, but not comparable to surface features (F-measure of 0.735). The segment length alone (measured in seconds) gives very good results, comparable to those of the complete feature set. As the full transcription is very time-consuming to carry out, this result gives us a good starting point to distinguish between patient and therapist and developing methods to judge the quality of the interaction between the two participants based on features from the audio only.

5.3. Quality Criteria Classification

As all sessions have to be rated on all quality criteria it makes sense – due to the limited data set size – to train models for the quality criteria independently. We use the same feature set we use for the distinction between patient and therapist also for the classification. Of the 35 sessions we have available to date, eleven have been segmented and partially transcribed. Table 1 indicates that the distribution of the classes is quite skewed and tends towards the higher marks. This supports our assumption that the data set might not be representative and contains very good therapists and motivated patients. Additionally, it indicates that the majority baseline is already quite hard to beat. As the classes 1-3 are hardly used, we collapsed them into one class, which leaves us with four classes to distinguish.

Table 5 shows the results for each quality criteria we looked into using ten-fold cross validation and a Random Forest learning algorithm. For most cases the classification is comparable or above the baseline. Only in the case of *Cognition* our results are below the baseline. These low results can be explained by the skewed data distribution, which often tends towards a rating of 5 or 6.

As the full transcription of the data is extremely time-consuming, we also experimented with the same machine-learning setup using only features available from segmenting the data[10]. Table 6 shows the results for these experiments. For most quality criteria the results drop, which is expected. Surprisingly, results for *Agenda* and *Behaviour* increased. While the results for both quality criteria did not exceed the baseline in the original setup, using the reduced feature set we achieve results better than the baseline.

Quality	Prec	Rec	F	Maj
Feedback	0.44	0.5	0.46	0.4
Cognition	0.31	0.42	0.34	0.5
Behaviour	0.25	0.17	0.2	0.35
Positive Focus	0.34	0.45	0.46	0.34
Agenda	0.39	0.5	0.43	0.43
Access Emotion	0.30	0.33	0.30	0.26
Cooperation	0.44	0.5	0.46	0.37

Table 5: Quality Criteria Classification Random Forest

Quality	Prec	Rec	F	Maj
Feedback	0.42	0.44	0.41	0.4
Cognition	0.25	0.21	0.23	0.5
Behaviour	0.5	0.46	0.48	0.35
Positive Focus	0.29	0.28	0.27	0.34
Agenda	0.46	0.52	0.47	0.43
Access Emotion	0.25	0.20	0.22	0.26
Cooperation	0.41	0.44	0.41	0.37

Table 6: Quality Criteria Classification Random Forest – segment features only

5.4. Most Important Features

In the following, we look at the individual quality criteria and also take a closer look at the best performing features. As we have some of the data fully transcribed, while some of the data was only segmented, we also looked at the best performing features in both feature sets.

Feedback The most important features to classify *Feedback* are social noise elements, but also the frequency of incomplete sentences. This indicates that both participants reflect to each other, that they are continiously listining and processing what the other is saying.

When we use only the segmentation based features, we observe that the duration of the segments is of imporantance, but also the relative frequency of segments by the therapist. As the therapist is using a lot of social noise, this increases the number of segments, although they are of course very short. Also very short breaks are of importance, indicating that there is little silence during the session.

Cognition The identification of *Cognition* is charactized by the importance of features such as the ratio of questions. It is interesting to note that also a high amount of stuttering and incomplete sentences is highly predictive features for this quality criterium. As is to be expected, words indicating sentiment or emotion are also very important.

[10] Please note that this is necessary in this setup, as we only received mono recordings of the sessions.

Based on the segmentation alone, we find that the breaks are very important in judging the quality of this criterium.

Behaviour The patients are required to also reflect on their behaviour. In our machine learning approach, we see that highly ranked features are those that are based on sentiment bearing words (both positive and negative), but not necessarily words associated with a specific emotion.
Using only segmentation features the amount of segments by the patient is very important. Additionally short breaks are mor important than longer breaks, contrary to classifying the quality criteria *Cognition*.

Positive Focus When determining the score for *Positive Focus* we see that the features ranked highest are those based on the LIWC dictionary. Interestingly, we see that a lot of negative emotions or sentiments are important and metaphoric expressions play a role.
When we use only segmentation based features, we see that short breaks are an important feature, but also the frequency of segments by the therapist.

Agenda Classifiying the quality criteria *Agenda* is also based heavily on words from the LIWC dictionary. Features based on words from the domain of space and time, but also sports are ranked highly. Also words from the domain of cause are important, which indicates that the participants reason about the why of their approach.
Using only segmentation-based features short breaks are important features for the classification.

Access Emotion For the classification of the category *Access to Emotion* we see that the duration of segments is of importance. Contrary to the description of this category, emotion or sentiment bearing words are less important in this machine learning based approach. Rather, surface features such as the number of characters per token and how many questions were asked are important.
The segmentation-based features also show a high interaction, as only short breaks play a role and the number of segments by the therapist are as important as the number of segments by the patient.

Cooperation The most important features for classifying the *Cooperation* category are also based on the LIWC dictionary. Among those are words from the category we, you and communication. This indicates a high amount of interaction and expressions that the two participants consider themselves as a team.
From the segmentation-based features short breaks and the duration of the segments are the most important features.

5.5. Discussion

Looking in detail at the results, we see that the full transcription of the data is necessary to achieve good results and that features based on the LIWC dictionary are especially important. These are only accessible through a thorough transcription. It is also interesting to note that for two quality criteria (*Agenda* and *Behaviour*) the smaller feature set based only on the segmentation features achieve considerably better results than using the full feature set. This is especially surprising for *Agenda*, where the top ranked features are actually from a matching domain in the LIWC dictionary. For *Behaviour* the features are less conclusive, with general sentiment features are ranked at the top.
It is also interesting to note that for both quality criteria that deal with the self-reflection of the patient among the high ranking features we see features that indicate difficulties such as hesitations, incomplete sentences and incomplete phrases. For most of the quality criteria the performance drops when only the limited feature set is used.

6. Conclusion and Future Work

We presented ongoing work in collecting and processing data from therapy sessions with patients with schizophrenic disorder in order to allow for semi-automatic processing based on natural language processing. Our aim is to classify the therapy quality, which is currently carried out manually with a high effort in time and expert man-power. Due to the recording quality a lot of effort as of now went into the transcription of the recordings. Based on the quality dimensions used by psychologists in the supervision of therapeutic sessions we identified specific dimensions which manifest themselves in natural language – either in *what* a person says or in *how* it is being said, and defined the extracted features accordingly. Our initial results indicate, that we can distinguish the therapist and the patient based on a range of features, most notably through information about the therapist segments only.
In the next step, we focused on the therapists and related their behaviour to the quality dimensions evaluated by experts, in order to build a classifier for these specific dimensions and evaluate them. Our results indicated that a high quality transcription is necessary to allow for a machine-learning based classification of the quality criteria used here. With the exception of two quality dimensions, which performed better with a reduced feature set, all quality dimensions were more reliably classified using the full feature set. Especially features based on the LIWC dictionary proved very valuable, which is in line with previous work.
In the future, we plan to extend the current feature set to also include acoustic features, which give an additional dimension of how people express themselves, beyond *what* they say, but rather *how* they say it.

7. Acknowledgements

This work was supported by the program *Forschung für die Praxis* by the Hessian Ministry for Sciences and Art, Germany and by the research center *Digital Communication and Media Innovation* (DKMI), University of Applied Sciences, Darmstadt. We specially thank our project partner at the Frankfurt University of Applied Sciences, under the supervision of Prof. Mehl who performed the manual quality annotation of the data and our two student helpers who were segmenting and transcribing the data.

8. Bibliographical References

Chakravarthula, S. N., Xiao, B., Imel, Z. E., Atkins, D. C., and Georgiou, P. G. (2015). Assessing empathy using static and dynamic behavior models based on therapist's language in addiction counseling. In *INTERSPEECH 2015, 16th Annual Conference of the Inter-*

national Speech Communication Association, Dresden, Germany, September 6-10, 2015, pages 668–672.

Flekova, L., Sousa, T., Mieskes, M., and Gurevych, I. (2015). Document-level school lesson quality classification based on german transcripts. *Journal for Language Technology and Computational Linguistics*, 30(1):99–124.

Görgen, W. and Engler, U. (2005). *Ambulante psychotherapeutische Versorgung von psychosekranken Menschen sowie älteren Menschen in Berlin*. Psychotherapeutenverlag – Verlagsgruppe Hüthig Jehle Rehm GbmH.

Hecht, R. M., Bar-Hillel, A., Tiomkin, S., Levi, H., Tsimhoni, O., and Tishby, N. (2015). Cognitive workload and vocabulary sparseness: theory and practice. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 3394–3398.

Heeman, P. A. and Allen, J. F. (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4).

Howes, C., Purver, M., and McCabe, R. (2013). Using Conversation Topics for Predicting Therapy Outcomes in Schizophrenia. *Biomedical Informatics Insights*, 6(1):39–50.

Kayi, E. S., Diab, M., Pauselli, L., Compton, M., and Coppersmith, G. (2017). Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 241–250, Vancouver, Canada, August 3–4, 2017. Association for Computational Linguistics.

Lecomte, T., Kingdon, D., and Munro-Clark, D. (2017). Cognitive therapy scale for psychosis – revised version. in preparation.

Lincoln, T. M., Rief, W., Westermann, S., Ziegler, M., Kesting, M.-L., Heibach, E., and Mehl, S. (2014). Who stays, who benefits? Predicting dropout and change in cognitive behaviour therapy for psychosis. *Psychiatry Research*, 216(2):198 – 205.

Lincoln, T. (2014). *Kognitive Verhaltenstherapie bei Schizophrenie*. Hogrefe, 2 edition.

Mehl, S. and Lincoln, T. (2014). *Therapie-Tools Psychosen*. Beltz Verlag, Weinheim Basel.

Meyer, C. M., Mieskes, M., Stab, C., and Gurevych, I. (2014). DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In Lamia Tounsi et al., editors, *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 105–109, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics. https://www.ukp.tu-darmstadt.de/software/dkpro-statistics/.

Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Paul, W., Alm, C. O., Bailey, R. J., Geigel, J., and Wang, L. (2015). Stressed out: what speech tells us about stress. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 3710–3714.

Schuller, B. and Batliner, A. (2013). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, Inc.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).