# What text mining analysis of psychotherapy records can tell us about therapy process and outcome

**Eleanor Rosa Yelland**

**UCL Division of Psychiatry**

**Supervisors: Professor Michael King, Professor Rumana Omar, Dr Ann Hayes, and Dr David Milward**

**A thesis submitted to University College London for the degree of Doctor of Philosophy**

**November 2016**

## Author’s declaration

I, Eleanor Rosa Yelland, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

01/11/2016

## Abstract

Increasing demand for mental health treatment and the transfer of a large portion of our lives online has led to the development of a growing range of computerized psychological therapy programmes. We are also creating and storing data at ever increasing rates, a trend that has led to the development of sophisticated textual analysis approaches.

This thesis sits at the cross-section of these evolving areas. It is an exploratory analysis of how text mining analysis can be applied to online cognitive behaviour therapy. The project emerged as a collaboration between two commercial partners: Ieso Digital Health and Linguamatics, and UCL. Ieso Digital Health provide online cognitive behaviour therapy via an online instant messaging platform and Linguamatics are the developers of text mining software I2E. The involvement of the two industrial partners in this project shaped two major components of this research; the data studied and the platform for textual analysis.

Linguistic analysis of textual data in mental health is a wide and variable field that brings together a variety of methods and data formats. These are broadly introduced in Chapter 1 and Chapter 2 provides a systematic review of research on the analysis of language used within therapeutic exchanges during mental health treatment. The research carried out in this thesis involved the development of a number of linguistic features within I2E and statistical analyses to explore their association with mental health outcomes and the development of predictive models of outcome. The results (Chapters 4-10) suggested that there were statistically significant associations between selected language features and therapy outcome scores but that these language features did not fare well as predictors of outcome when developed models were externally validated. These results and recommendations for the application of text mining in therapy transcripts are discussed in Chapter 11.

## Table of contents

## Table of figures

## Table of tables

# Chapter 1. Background

*This chapter introduces a number of background concepts to situate this research project within its clinical and research context. It provides information about the origins of the online therapy service provided by Ieso Digital Health and the patient population whose data will be used throughout the project. Text mining will also be introduced, the primary linguistic analysis method applied throughout this research. The final, larger part of this first chapter aims to provide the reader with an understanding of how language has been studied in the context of mental health research and practice.*

## 1.1 Mental Health context

### 1.1.1 Economic burden of mental health in the UK

Mental health services in the UK have been under ongoing and growing pressure for a number of years. The oft-reported statistic that one in four individuals will experience a mental health problem at some point in their life is enough to suggest that mental health is a major concern that cannot be disregarded or ignored. The Lansley report in 2011 brought together details of the numbers of individuals affected by mental health concerns and the economic costs associated with these. The report suggests that one in ten children between the ages of five and 16 have a diagnosable mental health disorder, most of which are likely to persist and worsen through to and in adulthood and nearly 50% of adults are expected to experience at least one episode of depression in their life. (HM G., 2011).

Associated with these high numbers are both the economic and personal costs of living with a mental health condition. Those living with a mental health condition may see an impact on their education and qualifications, their employment status and consequently income, and also socially with high levels of isolation (Corrigan & Rao, 2012; Rosenheck et al., 2006) as well as poorer physical health (Thornicroft, 2011). Beyond these personal costs, the 2011 report estimated the total cost of mental ill health to the NHS in England to be of approximately £77 billion, although this number could be closer to £105 billion. This first estimate accounts for loss of productivity and other work costs and payment for sickness and long-term absence associated with mental health as well as mental healthcare provision. Mental health accounts for 23% of the total burden of ill health yet only receives 11% of the health budget (The King’s Fund, 2015). Funding cuts made since then suggest that this may now be lower (McNicoll, 2015).

The 2011 report was part of an announcement of a government initiative to improve Mental Health services in England and raise its status to one of parity with physical health. One suggested step towards achieving this goal was to bring in the Improving Access to Psychological Therapies initiative (IAPT).

### 1.1.2 Mental Health in Primary Care and IAPT

In England, unless they are in an emergency situation, an individual who is suffering from a mental health issue is most likely to consult their GP in a first instance. In the case of anxiety or depression, a stepped care model is put forward by the NICE guidelines, meaning that different levels of treatment are offered depending on the severity of the problem (NICE, 2009, 2011). When an individual presents at their GP surgery asking for support they should first be offered psychological therapy in line with the severity of the problems they are experiencing. In the case of mild to moderate issues, patients are likely to be referred through IAPT. Other, specialist services may be recommended for those with more specific, severe or complex mental health issues, a complex trauma unit or specialist eating disorders service, for example.

The IAPT initiative was brought in to tackle low severity mental health issues, primarily depression and anxiety related conditions. Greater access to treatment was provided with the training of a large number of Psychological Wellbeing Practioners (PWPs) to work specifically within IAPT. These are graduate students who follow a cognitive behaviour therapy training course to deliver low intensity CBT initially and can later train as high-intensity therapists once they have gained adequate experience.

A main aim of the IAPT initiative is to provide short courses of CBT to those who are likely to benefit from it. This aims to reduce the numbers of people living with anxiety and depression without support and improve their ability to maintain or gain employment. The uptake was high, with over 1 million patients being referred for treatment with IAPT in the first three years and over 680,000 patients completing a course of treatment. The recovery rate has hovered around 45%, which is more or less in line with other treatments in mental health. Although the initiative originally only offered CBT, in the second phase, counselling, interpersonal therapy, couples therapy for depression and brief dynamic interpersonal psychotherapy have also been offered (Department of Health, 2012). Nevertheless, the majority of patients are referred for a brief course of CBT. Ieso Digital Health, a commercial partner of this research project, provides online CBT within the IAPT framework. Their caseload is therefore primarily made up of individuals dealing with a variety of mild and moderate depression and anxiety based mental health issues.

### 1.1.3 Ieso Digital Health service and caseload

The service provided by Ieso Digital Health mirrors that provided within NHS IAPT services. Patients are referred for treatment by their GP and allocated a therapist through Ieso. The first appointment with their therapist is an assessment session and the following sessions are treatment sessions. The number of sessions offered is dependent on the severity of the patient's difficulties. The difference with face-to-face treatment is that therapy sessions with Ieso are carried out entirely online, through a purpose-built instant messaging platform. Therapy sessions are carried out in real-time and patients have an online account through which they can make appointments, access their transcripts from therapy sessions, send and receive messages and files to and from their therapist and complete outcome score questionnaires.

The patients receiving treatment through Ieso Digital Health are referred with a large variety of mental health diagnoses and the service is currently expanding to be able to work with a wider pool of patients. The data to be studied within this project are primarily from patients referred with anxiety, depression or mixed anxiety and depression diagnoses. These made up the majority of Ieso Digital Health's caseload at the time of data collection.

#### 1.1.3.1 Depression

The term depression covers numerous diagnoses ranging from the more severe and chronic recurrent depressive disorder to the generally easier to treat single depressive episodes. It commonly refers to major depressive disorder (MDD). Depression-based conditions are characterized by low mood and a combination of a number of additional symptoms ranging from changing sleep patterns or eating habits to suicidal ideation and feelings of helplessness. These often have a considerable impact on an individual's everyday functioning.

The diagnostic criteria for first or single episode depression are much broader and it has been suggested that based on these criteria, up to 50% of the adult population is expected to experience an episode of depression at some stage of their life and that up to one in five individuals could be diagnosed as living with a depressive episode at any one time (R. C. Kessler et al., 2005). There are many subtypes of depression and both these and the severity that should lead to diagnosis are much debated within mental health research and practice (R. C. Kessler et al., 2010). For these reasons, variable rates of lifetime prevalence of depression are reported ranging from 6% (Weissman, Leaf, Florio, Holzer, & Livingston, 1991) to 25% (Lewinsohn, Rohde, Seeley, & Fischer, 1991). Despite the debate over specific numbers, there is no doubt that depression is a common problem throughout the population that can have a heavy impact on an individual's ability to maintain stability in their life on the social, educational or professional fronts.

#### 1.1.3.2 Anxiety

The term anxiety refers to another broad category of mental health problems that includes diagnoses such as generalized anxiety disorder, specific phobias, panic disorder and social anxiety. Given the variable definitions, prevalence estimates are again difficult to estimate accurately. Epidemiological work from 2009 estimates that between 6-12% of the population are affected by a diagnosable anxiety based condition (R. C. Kessler, Ruscio, Shear, & Wittchen, 2009). For certain specific diagnoses, the number can vary again with obsessive compulsive disorder or panic disorder suggested to affect 2% of the population, social phobia between 2% and 16% and generalized anxiety disorder between 3 and 30% (R. C. Kessler et al., 2009). Given the wide variations in these numbers and their sometimes contradictory nature it is difficult to obtain an accurate estimate of how many people are living with an anxiety-based mental health condition.

Anxiety-based conditions with more specific focuses are associated with similar symptoms but their onset is associated with a given worry such as social environments in social phobia or the fear of having a panic attack in public in agoraphobia. Anxiety is often co-morbid, especially with depression (Wittchen, Zhao, Kessler, & Eaton, 1994). Avoidance of particular situations is a typical behaviour in an individual with anxiety and if this then means an individual is missing out on social or professional opportunities, the path towards a low mood and loss of motivation is easy to trace (Tolman et al., 2009). Like depression, anxiety can be highly debilitating and is not necessarily a very visible condition. When professional help is sought, CBT is one of the recommended evidence-based treatments for anxiety-based conditions according the NICE guidelines.

### 1.1.4 Cognitive Behaviour Therapy for Anxiety and Depression

CBT has its origins in Cognitive Therapy for Depression, developed by Aaron Beck in the 1960s (A. T. Beck, 1979). It is based on the idea that there is constant interaction between thoughts, behaviours, physical feelings and emotions. This can create negative cycles, often termed vicious cycles, that maintain negative emotions and feelings. The aim of CBT is to understand the various elements that contribute to these cycles and find methods to change them and thus break the cycle. The first part of CBT will include formulation and acclimatizing a patient to breaking down situations into thoughts, emotions, physical reactions and behaviours. This may also involve considering a patient's past experience and exploring particular thoughts further to understand any core beliefs that an individual may hold. The treatment part of CBT will revolve around a therapist suggesting particular methods or techniques a patient might try out to break a vicious cycle (J. S. Beck, 2010). In the case of depression, this may involve worksheets where the aim is to generate alternate thoughts to interpret a situation to counterbalance automatic unhelpful thoughts. For example, if an individual with depression tried to call a friend and they did not answer the telephone, an unhelpful thought might be: 'My friend doesn't want to talk to me' whereas other options they would be asked to generate in a worksheet might be 'My friend is busy', or 'My friend is tired and wants some rest'. Each of these is likely to trigger a different set of emotions and potentially future actions. If a patient is able to recognize that an unhelpful thought that has no evidence to support it is less likely than an alternate, more realistic thought, then this may be a possible point at which to break the cycle. Behavioural exercises are also a central part of CBT, where patients are asked to put themselves in a particular situation or carry out a given action that they might previously have been uncomfortable doing and monitoring their and others' reactions. A variety of methods and exercises are used within cognitive behaviour therapy and these are selected by a therapist according to their patient's needs. The overall aim of CBT can be seen as providing the skills and techniques that may help an individual cope with mental health difficulties once they have been able to understand these.

There is a vast body of research looking at the effectiveness of CBT and at how it can be adapted to best suit the variety of mental health conditions that individuals present with. In the scope of this project, I am particularly interested in CBT for individuals with difficulties relating to anxiety and depression. What is the evidence around the effectiveness of CBT in these areas?

#### 1.1.4.1 Effectiveness of Cognitive Behaviour Therapy for Depression

Across a number of meta-analyses, CBT for depression has been found to be more effective than control conditions, where patients received no active treatment (Beltman, Voshaar, & Speckens, 2010; Cuijpers et al., 2010). When CBT is compared to other active treatments, however, the picture is more mixed. Three meta-analyses found CBT to be as effective as other psychological treatments (Beltman et al., 2010; Cuijpers et al., 2010; Pfeiffer, Heisler, Piette, Rogers, & Valenstein, 2011) but some individual studies found that CBT was more effective for depression than other psychological treatments (Di Giulio, 2010; Tolin, 2010). The success of CBT has also been compared to that of pharmacological therapy with both achieving a similar effect on chronic depressive symptoms (Vos T et al., 2004). CBT has also been considered to be useful in combination with pharmacological treatment when compared to CBT treatment alone. CBT for depression is generally considered to be an effective treatment option, a position that it is supported by its recommendation as the first line of treatment in the UK.

#### 1.1.4.2 Effectiveness of Cognitive Behaviour Therapy for Anxiety.

CBT is generally considered to be reliable in its effectiveness as a first-line treatment for anxiety-based mental health problems. As with depression, it seems that across different types of diagnoses, CBT is either equally efficacious or shows more improvement than other treatments. In the case of social anxiety, CBT showed a medium to large effect compared to control or waiting list with maintenance of the improvement at follow-up (Gil, Xavier, & Meca, 2001). In the case of generalized anxiety disorder, CBT was found to be effective compared to no treatment and placebo pharmaceutical treatment and equally as effective as relaxation, supportive therapy or pharmacological treatments. However, it was found to be less efficacious in comparison to attention placebos and in those with more severe diagnoses (Ruhmland & Margraf, 2001). A meta-analysis was carried out to look at particular elements or focuses of CBT such as exposure therapy, cognitive restructuring, development of social skills and considered both group and individual formats. Through comparison of effect sizes, the results suggested that these were equally efficacious (Powers, Sigmarsson, & Emmelkamp, 2008) and showed better long-term performance when compared to pharmacotherapy (Federoff & Taylor, 2007). Though this was a meta-analysis as opposed to direct comparison through clinical trial, it provides an indication of the presence of multiple active components to CBT.

### 1.1.5 Computerized Cognitive Behaviour Therapy and effectiveness

Improvements in technological understanding and the widespread access to the Internet and a variety of devices has seen the development of a number of different versions of CBT-based computer programs, computer-aided CBT or computer-enabled CBT. These range from online platforms that an individual navigates alone such as Beating The Blues, guided therapeutic programs where an individual completes a program of CBT modules alone but has regular review and the possibility of email exchanges with a therapist such as in SilverCloud, all the way to the service provided by Ieso Digital health where a patient and therapist meet in real-time on an instant messaging platform and the appointment follows the same structure one would expect in a face-to-face treatment session. A number of pieces of work and subsequent meta-analyses have been carried out to assess how well these online versions of CBT perform when compared to face-to-face treatment. A meta-analysis completed in 2009 found that the average effect size for computerised CBT for depression was 0.41, suggesting a moderate effect of the treatment as compared to control (Andersson & Cuijpers, 2009). There was, however, large variation in the effect size (Cohen’s *d*) when comparing therapies that were provided with support from a therapist or those provided without support from a therapist. This support could be provided over the phone, face-to-face or via e-mail. The average effect size for a computerized CBT course provided with therapist support was 0.61

whereas the average effect size of a CBT course provided without therapist support was 0.25. (Andersson & Cuijpers, 2009). Some earlier work found no difference between computerized CBT and that provided face-to-face in terms of effectiveness (Carlbring et al., 2005; Kiropoulos et al., 2008) but the study was not designed as an equivalence study between the two treatments so this would need to be done appropriately to draw robust conclusions. The authors of the meta-analysis note that the effect sizes associated with computerized CBT are in line with those of previous meta-analyses though they are a little lower. They are, however, not lower than effect sizes associated with psychological treatments provided in primary care (Cuijpers et al., 2009).

In the case of computerized CBT for anxiety, a meta-analysis carried out in 2009 found that the effect size comparing computer-aided CBT and non computer-aided CBT was -0.06 (95% CI: [-0.22 ; 0.10]) suggesting that there was no significant difference in the outcomes associated with computerised CBT as compared to face-to-face CBT (Cuijpers et al., 2009).

As mentioned above, ‘computerised CBT’ is a term that can cover a wide range of therapy provision in terms of format and amount of support provided. The specific therapy format that this project is working on was the subject of a clinical trial published in 2009. The study compared therapist-delivered online CBT in addition to care as usual and care as usual whilst being on an 8-month waiting list for online CBT. The results suggested that in the treatment group, 38% of participants had recovered from depression at a four-month follow-up whereas this number was only 24% in the control group, suggesting that this form of online CBT was effective. The results suggested an effect size of 0.81 associated with the therapy intervention (D. Kessler et al., 2009).

## 1.2 Linguistic analysis and application in Mental Health

This research project looks at the language used by patients and therapists during their therapy sessions. The aim is to explore and analyse the language so as to determine what can be learned about the therapeutic process and mental health outcomes.

The speed gained from computerising quantitative methods of linguistic analysis and an increased access to large textual corpora online mean that the potential benefits of analysing language within mental health are very much a current topic (Coppersmith, Dredze, Harman, & Hollingshead, 2015; Mitchell, Hollingshead, & Coppersmith, 2015). Linguistic analysis has been put forward as a method of illustrating individual and group differences in mental health status both in parallel and over time (Cohn, Mehl, & Pennebaker, 2004; Rude, Gortner, & Pennebaker, 2004), as a validation tool for psychometric scales (Tov, Ng, Lin, & Qiu, 2013) or as a method of monitoring progress in treatment (Arntz, Hawke, Bamelis, Spinhoven, & Molendijk, 2012). A number of groups have looked to the use of language as a window into the mind and thus consider it an opportunity to gain further understanding of mental processes and amongst other areas, mental ill health. Section 1.3 provides a broad review of the linguistic analysis research methods that have been applied within the mental health field where the approaches introduced above will be expanded upon. A systematic review that focuses specifically on linguistic analysis within therapeutic dialogue will make up the second chapter of this thesis.

### 1.2.1 Linguistic analysis and corpus linguistics

Linguistic analysis refers to the scientific study of language and is often seen to cover 5 broad areas: phonology, morphology, syntax, semantics and pragmatics. In this project the focus will primarily be on syntax and semantics, respectively, the grammatical structure of the language under study and the vocabulary or words that are used. Linguistic analysis includes a wide variety of methods and this piece of work sits within the area of corpus linguistic research. Corpus linguistics was developed alongside empirical research work as a tool to explore, develop or test hypotheses by looking at the variations in language features. Corpus linguistics encompasses both corpus-driven and corpus-based linguistics. In corpus-based linguistic study linguistic features are defined prior to the analysis and it involves analysing the variation and the use of these features in a given set of textual data. In corpus-driven linguistics, the linguistic features are not strictly defined prior to working on the text but are considered to emerge from the analysis of a corpus (Biber, 2009). This can be through both qualitative and quantitative methods or a mixed methods approach.

There has long been interest in the words people use and what they might imply about their mental state, Freud's free association work is a famous example of this. It is only recently, however that the application of quantitative analysis methods to linguistic data in mental health has really emerged and grown very rapidly. Within its field, linguistic analysis is often applied to learn about the language itself and how its uses change and evolve. Within fields such as sociolinguistics or linguistic analysis in mental health it is being considered more in terms of a tool that can provide a new perspective or approach to known or current issues. For example, looking at the natural language an individual uses to describe and understand their experience with a mental health disorder or their treatment may provide insights that surveys or symptom measures do not capture. In this project, I will be using specialized text mining software to develop and extract linguistic features. The association between these and outcome scores will then be explored prior to building predictive models of outcome. The long-term aim is to improve the understanding of therapy process and service provision. For example, if analysis of the language used in therapy can provide information about whether a patient is likely to successfully complete their course of treatment or suggest particular features of language that are associated with good or poor outcomes, then this can be used by the service provider to recognize patients who may need extra support or alternately, who are not suited to this particular therapy format.

### 1.2.2 Text mining with I2E by Linguamatics

There is a large range of methods and approaches to linguistic analysis, some of which will be covered in the literature review in the next section 1.3.

Given the nature of this project as a collaboration with industry and the focus on the potential for text mining using I2E by Linguamatics, a specific approach to language analysis with a given software, it is important to provide some detail about this method at this stage.

Text mining emerges from the field of information extraction. It aims to derive or extract relevant and high-quality information from textual data. It involves the development of search phrases, referred to as queries, through which large amounts of text can be searched and relevant results returned. More refined than a simple keyword search, text mining involves the application of natural language processing techniques to facilitate the search and data extraction. This includes natural language processing algorithms for different stages of natural language processing. These are processes such as parsing, also know as parts of speech tagging, which is the process by which the string of characters making up the text is split into its component parts and the grammatical role of each phrase and word is determined. The syntactic structure of the text is automatically detected and individual words and phrases are labeled as verbs, nouns, adjectives and so on. Grammatical, syntactical or morphological rules can be incorporated into a search query in order to extract concepts that, for example, appear within the same phrase or with a verbal relation linking them.

Within I2E, dictionaries, referred to as ontologies, can be developed for specific fields or purposes. This allows users to search for complete lists of key terms and manually develop queries through a user interface that allows the visual representation of a given query. The query development process is iterative and subjective in that after an initial build, a query is run and the results then evaluated to determine whether it is providing relevant results. Where this is not the case and there is space for improvement of the query, it is then edited and resubmitted to evaluate the changes, and hopefully improvements, in the returned results. Further detail on the query development process will be provided in section 3.3.1 in the context of individual queries.

Text mining differs from word count methods in that the primary focus is on extracting particular concepts and phrases, rather than broader linguistic categories, although this kind of work can also be carried out with I2E. The more content-focused approach, however, suggests that the researcher needs to have an idea of what they are looking for in text prior to embarking upon the search as query development generally relies on an iterative process of query building, checking results and editing the query as required. Text mining has been applied in the context of drug discovery as it allows large amounts of text to be searched through with relative ease and bring up indirect links between compounds and symptoms (Milward, Blaschke & Neefs, 2006). Within health research it has been used to assist with pneumonia diagnoses from unstructured text in radiology reports (Liu, Clark, Mendoza et al., 2013).

To the author's knowledge, text mining has not previously been applied in transcripts from mental health treatment. There is therefore no precedent on which to base this work. A range of work has, however, been carried out looking at a variety of methods of language analysis in mental health research. This body of work will be described in the following section to provide some background into the relationships between features of language and measures of mental health state. This will not be an exhaustive review but aims to provide an overview of the type of research that has been carried out, a more focused review on computerized analysis of language within psychological therapy can be found in the next chapter.

## 1.3 How have computerized methods of linguistic analysis been applied to mental health research?

To go about answering this question a literature search was run on three databases: Web of Science, Medline and PsychInfo. Though this selection is mainly focused on health, PsychInfo does include Linguistics-focused papers and Web of Science has a broader coverage with topics reaching across the sciences and humanities. Nevertheless, it is possible that the inclusion of databases such with a specific focus on linguistics and computational methods (such as NLP) would have improved the reach of the literature search. These could be databases such as Linguistics and Language Behaviour or ArXiv, a database of preprints of journal articles across fields including mathematics, computer science, quantitative biology and statistics amongst others. Two sets of search terms were developed, the first covering terms relating to mental health and the second, terms relating to automatic analysis and linguistic analysis. The final search terms included were as follows: (“text mining”, “natural language processing”, “information retrieval”, “linguistic analysis”, “sentiment analysis”, “computeri?ed analysis”, “data extraction”, “textual data”, “pattern recognition”, AND, “Psychol*”, “Psychiatr*”, “Clinical Psych*”, “Cognitive behaviour* therapy”, Mental Health[MeSH Major Topic]). Results from this literature search were supplemented through hand searching of references in relevant articles. This search and review was not run as a systematic review but aimed to provide an understanding of the research background to this thesis.

The literature search found no recent review covering the applications and potential of automatic language analysis in psychopathology and treatment. Tausczik and Pennebaker do mention some previous work in the context of the development of the Linguistic Inquiry and Word Count software but this is by no means exhaustive (Tausczik & Pennebaker, 2010). Prior to this, a review by Garfield et al., published in 1992 looked at the potential of artificial intelligence (AI) and natural language processing in psychiatric research and treatment. Garfield introduces the theoretical concepts that are the basis of the methods in the field such as looking at morphology, syntax, semantics and pragmatic elements in language (Garfield, Rapp, & Evens, 1992). Two AI systems, ELIZA (Weizenbaum, 1966) and PARRY (Colby, Weber, & Hilf, 1971) were designed to understand and respond to language simulating a Rogerian psychotherapist in the case of ELIZA and a paranoid patient in the case of PARRY. Although not strictly based on automated linguistic analysis, it is important to note that both of these systems used pattern matching, a method that is likely to be of great relevance to computerized linguistic analysis. Pattern matching involves picking up common structures, often revolving around a verb, and looking at the words that are associated them (e.g. ‘X ate Y’ informs us of a particular relationship between X and Y). Garfield also reports on the General Inquirer system. This is a method that is more closely associated with the computerised linguistic analysis we are considering in this project. It involves mapping counts of word categories onto a given text as defined by the attached dictionary. This method is also the only computerised method that Tausczik and Pennebaker (2010) report on when introducing the Linguistic Inquiry and Word Count software and method.

The published literature on computerised linguistic analysis in mental health forms a varied and incomplete picture of its potential. The research carried out mostly falls into four broad categories. These are detailed below with a brief explanation of the method itself and an exploration of what has been achieved with its application in the area of mental health research.

### 1.3.1 Word Count and dictionary-based methods

#### 1.3.1.1 General Inquirer

The General Inquirer was an early attempt to create an automated content analysis system with the aim of discovering and extracting psychological themes from the language used in group discussions. It was originally developed in 1962 and relies on a dictionary made up of 164 categories that include the 3000 most frequent words in the English Language (as determined by (Thorndike & Lorge, 1944)) and additional sets of words that were deemed by the authors to be relevant to the context of behavioural science (Stone, Bales, Namenwirth, & Ogilvie, 1962). As is the case for most language analysis systems currently available, the system also incorporates syntactic information to assist analysis. The syntactic function of the language used is considered alongside semantic meaning. The General Inquirer has been applied by a couple of research groups in the context of the language used by individuals with a diagnosis of schizophrenia. Maher, McKean and McLaughlin in 1966, conducted a series of analyses on over 100 texts written by hospital in-patients in order to explore features of language use that would characterize symptoms of schizophrenia(Maher, McKean, & McLaughlin, 1966a). After an exploration of the data with the General Inquirer, they hypothesised that a high object to subject ratio in patient language was seen as indicative of thought disorder, a common symptom of schizophrenia. This hypothesis was justified by the idea that generally speaking, a sentence in English contains one object to each subject. The presence of more than one object, and thus a high object to subject ratio, puts forward a grammatically disorganized sentence, e.g, 'I went to the shop and the bank and the post office.' And was seen as evidence of thought disorder. This is an example of language being looked at as holding potential evidence of symptoms in its structure and style as opposed to the content of their speech. Despite support for their hypothesis across two initial sets of analyses, on the third replication the idea of the object-subject ratio as a discriminatory factor between groups of individuals with and without a diagnosis of schizophrenia was not supported (Maher, McKean, & McLaughlin, 1966b).

In 1975, the system was applied to a collection of speech samples and dream transcripts from individuals with a diagnosis of schizophrenia and control individuals (Tucker & Rosenberg, 1975). This analysis found that the two groups were differentiated on 14 of 84 categories considered. The results suggested that patient speech translated an individual's struggle to place themselves in time and space as well as an attempt to cope with confusion and internal psychological discomfort (Tucker & Rosenberg, 1975). Replication a year later with a larger sample of individuals with and without a diagnosis of schizophrenia found that only 3 of the 31 tested categories showed significant differences between the clinical and non-clinical groups thus failing to support the 14 categories that had previously discriminated between clinical and non-clinical patients. The three categories in this case were: negations (e.g 'not', 'don't'), with a high proportion of these in the clinical sample, and pleasure and ascent-themed language ('improve', 'go up') that were both underrepresented in the clinical sample (Rosenberg & Tucker, 1976). The results in these pieces of work do not seem stable and the explanations that attempt to ground them in theory are not robust. It seems that the factors that appeared to discriminate best between clinical and non-clinical groups were not those hypothesized, again suggesting the possibility that these were chance findings that require replication.

#### 1.3.1.2 Linguistic Inquiry and Word Count

##### 1.3.1.2.1 Description

In recent years, the most popular word count software for language analysis within the context of psychological research has been the Linguistic Inquiry and Word Count (LIWC), developed by Pennebaker, Booth and Francis, originally in 2001, with updated versions published in 2007 and 2015. It contains a dictionary of over 3,500 words sorted into over 80 categories (Pennebaker, Booth, & Francis, 2007). The LIWC allows the user to input a block of text and provides an output spreadsheet containing frequency measures for each category within the text. These are generally presented as a percentage of the total number of words in the text.

The software is simple to use and requires little technical understanding. It is therefore seen as a good option to obtain some quantitative measures from data more often approached from a qualitative perspective. The method was proposed and developed on the assumption that the words an individual uses 'convey psychological information over and above their literal meaning and independent of their semantic context' (Pennebaker, Mehl, & Niederhoffer, 2003). It has been criticized for this very point; for being subject to a number of ambiguities in language as it lacks the capacity to take context into account (Bantum & Owen, 2009). Given that the output from the LIWC analysis is a percentage of words from each category in the text, it can be considered a crude method of analysis that does not exploit the full potential of the textual data being studied. Nonetheless, it has been shown to detect significant differences in language use across gender, age groups and personality, for example, and to provide useful insight into understanding how word choice and use are associated with aspects of an individual’s character and mental health (Pennebaker et al., 2003).The 2003 review paper by Pennebaker and colleagues covers many areas in which significant differences in language use have been hypothesised and examines the evidence supporting them. For example, language use appears to be a subtle marker of age with higher levels of positive language, and lower levels of negative language in an older population sample (Pennebaker & Stone, 2003). Gender differences appear to be inconsistent, however (Pennebaker, Mehl, & Niederhoffer, 2003).

A 2005 paper looked into construct and concurrent validity of the LIWC in the analysis of messages in online support groups for women with breast cancer diagnoses (Alpers et al., 2005). For this, the LIWC scores obtained from the analysis of 521 messages in an online support group for individuals with breast cancer were compared to scores obtained in emotional and unemotional writing in a previous study (Pennebaker, Booth, & Francis, 2001) and breast cancer related newspaper articles. The results suggested that the LIWC profile of the language used in the online support group was more similar to the emotional than unemotional writing and to an article about the experience of a cancer diagnosis rather than an article about it’s genetic markers. These results were put forward as supporting the use of the LIWC as a method of analyzing language in online support groups (Alpers et al., 2005). Similarly, LIWC scores were compared with results from human ratings and found to be moderately correlated (Alpers et al., 2005). However, the human rated categories were not defined to directly match the LIWC categories. For example, it was expected that the ‘Body’ category in the LIWC would be correlated with a human rated category defined as ‘Medical Aspects of Cancer’. This was the case but the categories were not developed to be directly equivalent, making it difficult to determine what an ‘ideal’ correlation supporting validity of the LIWC category would be.

Two further pieces of work looked into validation and reliability of the LIWC. The first compared LIWC category measures in a variety of corpora including emotional personal narratives, technical scientific writing and fictional novels in order to determine that the majority of words used were being measured and that the different categories of language were achieving significantly different scores in LIWC categories. This was found to be the case and that 86% of words used were being measured (Pennebaker et al., 2001). Scores for self-report, human ratings and LIWC measures were compared for text samples describing an everyday object or the experience of going to university. These included emotional and cognitive dimension selected to match LIWC categories. High correlations between the self-report, human ratings and LIWC categories support the use of the LIWC as measure of emotion in language (Pennebaker et al., 2001). It seems that the LIWC is a straightforward and easy-to-learn method of approaching natural language for analysis. Word frequency is a relatively crude method of measurement, meaning that full exploration of the textual data is limited as compared to qualitative methods or more complex computerized linguistic analysis. Nonetheless, it has successfully been applied to detect differences between groups and change in individuals as further detailed below.

Ramirez-Esparza and colleagues suggested a Spanish version of the LIWC by directly translating words that make up the dictionary (Ramirez-Esparza et al., 2007). The dictionary has also been translated into a number of other languages: Chinese, Arabic, Dutch, French, German, Italian, Portuguese Russian, Serbian and Turkish (Alparone, Caso, Agosti, & Rellini, 2004; Bjekic, Lazarevic, Zivanovic, & Knezevic, 2014; Hayeri, Chung, Booth, & Pennebaker, 2010; Kailer & Chung, 2010; Piolat, Booth, Chung, Davids, & Pennebaker, 2011; Wolf, Sedway, Bulik, & Kordy, 2007; Zijlstra, van Meerveld, van Middendorp, Pennebaker, & Geenen, 2004).

The LIWC approach has been applied to the analysis of between group differences comparing mental health patients and non-clinical control participants, as well as within group differences over time.  In addition a number of projects have focused on how LIWC measures relate to established mental health scales. I will initially report on research into between group differences in language use as measured by the LIWC.

Prior to discussing the literature in the next few paragraphs, it is useful to define the concepts of positive and negative language as these recur frequently both throughout the literature and this thesis. Sometimes called ‘affective’ language and also called sentiment language, the concept of positive and negative is familiar but defining what is considered positive or negative can be more difficult. More specifically, it is, as with most language-related aspects, context-dependent. In the context of this project, positive language refers to words and phrases an individual may use to express pleasure, happiness, confidence or success, for example. Negative language will cover areas such as anger, anxiety, sadness or disappointment. With most word-based dictionaries, the relevance of individual words is determined by human judgment, generally requiring the agreement of 3 or more raters.

#### 1.3.1.2.2 Group differences

The most common and recurrent finding in computerised applications of LIWC analysis is the higher percentage of first person singular pronouns and negative language words used in groups of individuals presenting with a depressive or anxiety-based condition as compared to a control group. The stronger and more reliable effect is present in the context of major depressive disorder. This has been noted in a number of expressive writing research projects that focused on disclosure as a therapeutic exercise (Pennebaker et al., 2003a). Disclosure in this context refers to the process of talking or writing specifically about an event, topic or situation and the personal thoughts and feelings associated with it. A 2008 study applied both the Spanish and English versions of the LIWC to posts on online forums written by depressed and non-depressed individuals in Spanish and in English and found similar results across the two languages that support these previous results (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker, 2008). Evidence from language used on Twitter supports the evidence for higher levels of negative emotion (specifically anger) in language used by those who identify as depressed (Park, McDonald, & Cha, 2013). This effect is further supported by evidence from writing samples from students who had previously been diagnosed with clinical depression as compared to a control group (Rude et al., 2004) as well as in a comparison of journals written by anorexia nervosa patients and control participants (Wolf et al., 2007).

Conversely, a 2010 paper by Molendijk et al., reports on an unsuccessful attempt to replicate the findings of Rude et al., (2004). The participant sample was made up of individuals with a personality disorder diagnosis who were divided into groups by whether they were currently experiencing, had previously, or had never experienced depression. They found that instead of the differences in negative language use and first person pronouns being between the depression-based groups, these differences were found across the psychiatric group in comparison to a non-psychiatric control group (Molendijk et al., 2010). A recent comparison of language contained in autobiographical memories in groups made up of individuals with major depressive disorder (MDD), borderline personality disorder (BPD) and a control group suggested that the individuals with borderline personality disorder used more first person pronouns than the control group but not the MDD group, supporting the work by Molendijk et al., (2010) but that the BPD group used more anger and social words than both the MDD and control groups (Rosenbach & Renneberg, 2015).

Evidence of group differences in the language used can also be found across other mental health issues. A comparison of written samples by individuals with and without a history of child sexual abuse (CSA) also showed differences in first person singular pronouns between the two groups (Lorenz & Meston, 2012). The participants were asked to write two personal essays, one neutral about the previous 24 hours of their life and one asking them to write down their deepest thoughts and feelings about sex and sexuality. The group of individuals with a history of CSA were found to use higher levels of first personal singular pronouns when writing about sexual topics as compared with the group without a history of CSA (Lorenz & Meston, 2012). This result could be seen as evidence of the lasting effects on adult mental health of sexual abuse in childhood. A slightly differing result was found in a group of individuals with a social anxiety disorder when compared to a non-clinical control group. Here it was not the general category of negative language that was heavily represented, but specifically fear and anxiety related language (B. Anderson, Goldin, Kurita, & Gross, 2008). For this study, participants had been requested to recall autobiographical memories that were specifically salient in terms of humiliation, shame or embarrassment. The results illustrate how emotions in the author can translate and be quantified in the language they use to express themselves and specifically, that the LIWC is able to identify these.

The presence of increased levels of first person singular pronouns in individuals suffering from depression or an anxiety-based mental health disorder is consistent with self-focused theories of these difficulties (Mor & Winquist, 2002). There is a suggestion that certain aspects of these mental health disorders lead to an increased self-focus. In the case of depression, rumination or negative automatic thoughts may lead an individual to be more focused on themselves. In the case of anxiety based mental health issues, worry focusing on the self or the consequences of actions or situations on the self as well as worry about their ability to cope may be a source of self-focus. The presence of quantifiable differences in personal pronoun use and the suggested association with aspects of mental ill health leads to the suggestion of using text-based screening tools. These could be designed to be sensitive to abnormally high levels of both first person singular pronouns and negative language in the language used by an individual. This could also be a method of looking at symptoms and behaviours that requires no additional input from a patient other than a language sample.

Results from a research project by Junghaenel and colleagues serve as a reminder that this effect is not necessarily present across all mental health difficulties (Junghaenel, Smyth, & Santner, 2008). Their work involved the comparison of writing samples from psychiatric outpatients and non-clinical controls. The hypothesis that significant differences in first person pronoun use suggest a link between high levels of personal pronoun use and high neuroticism (Pennebaker & King, 1999) was not supported. This contrasts with results from Molendijk et al. (2010) who saw an effect across individuals with a personality disorder diagnosis that was not restricted to individuals with a history of or current depression. It is important to note, however, that the sample in Junghaenel et al., (2008) contained a large variety of diagnoses and was not specifically focused on depressive disorders. Moreover, other differences in linguistic measures were found, namely in the frequency of optimism and energy related words and in measures of cognitive processing language where psychiatric patients used less discrepancy, inhibition and tentativeness language (Junghaenel et al., 2008). Discrepancy words are words such as ‘should’, ‘would’, ‘could’, the inhibition category includes terms such as ‘block’ or ‘constrain’ and tentativeness language includes words such as ‘maybe’, ‘perhaps’ or ‘guess’. These results may seem counter-intuitive as we might expect expressions of uncertainty and discrepancy to be more frequent in a clinical group. However, this is an example where the linguistic differences between groups are perhaps more subtle and thus, more informative within a given context. Rather than providing an overarching effect that may be associated with a mental health problem in general, these results may be illustrating specific differences between the groups in this study. It is also likely that effects are highly context-specific with the format of writing, instructions provided to participants and circumstances of those writing, all being important. In this case participants were recounting an important event in their life and it was suggested that the greater use of tentative, discrepancy and inhibition language in the non-clinical group was associated with more in depth consideration and qualification of the events as opposed to a more straightforward recounting of them that occurred more frequently in the psychiatric group (Junghaenel et al., 2008). These results were, however, still exploratory and potentially unreliable given the small sample size used for the study with only 17 participants in each group.

There have been a number of studies of differences in language use in those living with an eating disorder. A 2006 study by Lyons and colleagues picked up on the differences between ‘pro-anorexia’ online message board postings and those from individuals in recovery from anorexia. ‘Pro-anorexia’ websites are highly controversial websites that appear to promote anorexia and are a platform for discussion and sharing experiences. A number of campaigns have looked to ban and shut down these websites and limit or remove any links to them on social media. The language used in these message board postings was suggested to contain more positive language, less anxiety language, less cognitive reflection and less self-directed attention. These results suggest the difficult idea that this group is attached to anorexia despite its devastating physical effects, possibly as a form of coping (Lyons, Mehl, & Pennebaker, 2006). These results were partially supported by results from Wolf, Theis & Kordy (2013), who looked at the language in a number of blogs that were either pro-anorexia, blogs following an individual’s recovery journey or control blogs not associated with eating disorders. The authors applied LIWC analysis and found more closed-minded (i.e high self-focus with few social references) language, less negative emotion language and more food-related language in the pro-anorexia blogs as compared to recovery journey blogs (Wolf, Theis, & Kordy, 2013). The results relating to self-focus were opposed in these two studies but the results for emotional language were similar. The negativity associated with recovery may provide some insight into the level of treatment resistance commonly associated with anorexia. Working towards recovery can be extremely emotionally challenging and this may be what is reflected in the language studied.

A second piece of work carried out in 2013 looked at LIWC measures of language in a group of individuals with a diagnosis of anorexia nervosa (AN) and a control group and found higher levels of negative language in the AN group but also found that the body mass index (BMI) was positively correlated with negative language in the clinical group and not the control group (Brockmeyer et al., 2012). This work suggests that higher BMI was associated with higher negativity, but only in the group of individuals with anorexia. Other work around eating disorders focused on the impact of disclosure tasks and compared the language use and clinical outcomes of a group of young female participants in a partial-hospitalisation programme for eating disorders who were assigned either to a traumatic disclosure or control writing task. No significant difference in clinical outcomes were found but there were quantifiable differences in the language use such as increased negative, cognitive, and insight language and function words in the traumatic disclosure group (Gamber, Lane-Loney, & Levine, 2013). These results may not be surprising given the writing task allocated to each group and the study has relatively low power with only 21 participants, it nonetheless provides evidence of the LIWC's ability to pick up differences in narrative focus.

Further examples of group differences are as follows. In the case of a comparison between two groups who were new or long-term residents in a mental health facility, lower levels of negative, metaphysical and cognitive processing language were found in the group that had been in the facility longer (Saavedra, 2010). These differences in language could be seen as evidence of positive adaptation to the environment or resignation to remain within the mental health facility. An earlier study working with adult survivors of child sexual abuse in comparison with a control group looked into hyposexuality. It was noted that when writing about a sexual topic, the adult survivors used more negative language than the control groups and this was significantly correlated with their scores on hyposexuality scales (Rellini & Meston, 2007). Finally, Handelman & Lester applied LIWC methods to a set of suicide notes from individuals who either attempted or died by suicide. This is a dataset that will be referred to again as it has been studied using a variety of methodologies. In this study, only five small, yet significant, differences were found across all the LIWC categories with more second person pronouns and 'hearing' language ('listen', 'hear'), references to other people and use of the future tense in notes from individuals who died by suicide and fewer terms relating to inclusiveness and metaphysical topics (Handelman & Lester, 2007). The sample groups were, however, not balanced and the study had low power and a high risk for confounding variables.

The variety in the work described above shows the diversity in topics and populations that have been under study with the LIWC and the information that can be gleaned from this type of analysis. The ability to quantify aspects of language allows for new insight into the writers' thinking processes and motivation as well as providing a platform for comparison and measurement. However, these examples are just as much a reminder of how important interpretation and an understanding of the context is and that great care should be taken when drawing conclusions in this young and developing field.

#### 1.3.1.2.3 Changes in language use over time

*1.3.1.2.3.1 Therapeutic interventions and experimental conditions*

This section will cover work that considers changes over time within the same group, often in addition to group differences. A first example looks at language in personality disorders. Essays collected at three time-points from individuals with personality disorders undergoing psychotherapy showed significant changes in language use when analysed with LIWC software (Arntz et al., 2012). The results suggested that use of positive language increased and use of negative language decreased in the clinical population. They also compared results with language use in a non-clinical control group and found that the originally significant difference in use of personal pronouns between the two groups gradually shrank at each time point. Overall, the characteristics of language use in the clinical group approached those of the non-clinical group as treatment progressed, even if some differences remained. In a separate study involving older adults completing disclosure tasks it was found that improvements in depressive and physical health symptoms were predicted by a decrease in first person singular pronoun use and sadness language and increases in insight and causal words (Consedine, Krivoshekova, & Magai, 2012). This particular piece of research focused on the benefits of disclosure in different writing conditions (i.e writing about a sad event in a positive or neutral condition). Both pieces of work described here are in line with previous research in suggesting that increased first person singular pronoun use was indicative of affective problems or anxiety and that measuring positive and negative language use can potentially provide insight into an individual's progress in treatment.

In keeping with this idea are the results of a project that put linguistic analysis forward as a method to assess the effect of a mindfulness intervention in a therapeutic community for substance use recovery as compared to treatment as usual in this community (Liehr et al., 2010). Though the primary goal of the research was to assess the intervention, it is the use of language analysis as a measure of change that is of interest here. The trial was not designed as a randomised trial but used data from previous patients as control data and patients newly enrolling were all offered the mindfulness intervention. The results showed a decrease in anxiety and negative language and an increase in positive language use, irrespective of group status. These changes were generally indicative of improvement, suggesting that linguistic analysis can be considered as a method of monitoring and observing therapeutic change and psychological state. Supporting this idea were results from a piece of work looking at language before and after treatment for female survivors of child sexual abuse. They found evidence that a reduction in first person singular pronoun use and an increase in positive language was associated with reduced depression symptoms (Pulverman, Lorenz, & Meston, 2015). Beyond emotional language, Arntz and colleagues highlighted with their work (mentioned previously) the heavy presence of negations in individuals with personality disorders, suggesting that patients often focus on what they miss or don't have in their lives, which could be a source of distress and an area worth addressing in psychotherapy

(Arntz et al., 2012). This study is an example of how linguistic analysis of writing samples could help inform and improve treatment.

Two recent studies have considered how language analysis can inform our understanding and practice of web-based psychological treatment. The first was an online group therapy setting aiming to work on mood ‘mastery’ and looked at the relationships between language at baseline and throughout the course of therapy, and adherence and outcome. The results suggested that fewer negative and more discrepancy words at baseline were associated with higher mastery at baseline (lower severity) and less discrepancy language and more social words were associated with better adherence (Van der Zanden et al., 2014). Discrepancy language includes terms such as ‘should’, ‘would’ and ‘could’ that suggest incompatibility or dissonance between two elements, in this context most likely between what an individual thinks they should be and what they are doing. There was also a significant correlation between changes in discrepancy language over the course of therapy and changes in depression scores with an increase in discrepancy being associated with a decrease in depression score (Van der Zanden et al., 2014).

The second piece of work follows the changes in the language contained in patient to therapist communications over a course of therapist assisted internet CBT. This intervention consists of a number of modules that an individual primarily works through alone, although they are in contact with a therapist who will respond to them weekly. It is the language contained in the messages sent to the therapist that was analysed here. It was found that over the course of the intervention, rates of negative language, anxiety, causation and insight words reduced, further supporting the body of evidence described here (Dirkse, Hadjistavropoulos, Hesser & Barak, 2015).

*Non-experimental conditions*

In the case of non-experimental conditions, such as traumatic events or learning of a serious health condition, for example, analysis of language has also provided some interesting insight into the psychological changes that might occur in the individual over time. Changes in verbal behaviour following the events of September 11th 2001 were investigated by two research groups. Cohn and colleagues looked at markers of psychological changes in 71,800 online blog posts from 1,084 people before and after September 11th (Cohn et al., 2004) and D'Andrea et al. looked at writing samples from undergraduate students in the Boston area immediately following, and 6 months after the events, in order to examine linguistic predictors of PTSD symptoms as measured by the Impact of Event Scale-Revised (Weiss & Marmar, 1996), (D'Andrea, Chiu, Casas, & Deldin, 2012). In the first study, positive emotion levels dropped, and cognitive processing, social orientation and psychological distancing increased after the events, as compared to baseline. Cognitive processing is a LIWC category that includes words such as 'think' and 'question' and aims to measure participant understanding of the issues they address. Social orientation considers terms such as 'talk', 'share', and 'friends' to determine how much an individual is focused on their social world and psychological distancing was defined within this study as a variable made by combining LIWC measures. It combined measures of articles, words of more than six letters, pronoun use, discrepancy language and present-tense verbs. It aimed to distinguish between individuals focusing on the personal and the here and now and those using a more abstract and rational tone. Though it isn't clear from the report, the measure of social orientation appears to be obtained from the established LIWC category of social processes, supporting its validity as a measure here. The interpretation of such a word-based category will be context-dependent and must be done with a good understanding of how it is measured. The concept of psychological distancing, however, is more complex and was created by this group. It does not appear to have been independently validated in the same way the individual LIWC categories were. The measure combines a number of LIWC linguistic categories such as use of words longer than 6 letters or inverse scores of first singular pronouns and discrepancy. These features are reported to correlate in natural language and are put forward as markers of an 'abstract, impersonal and rational tone' but independent validation of the psychological distancing measure would strengthen the results put forward in this paper.

The results suggest that positive emotion language returned to baseline approximately a week after the events. Other measures did so after 2 weeks but social orientation dropped below baseline in the 3-8 weeks after September 11$^{th}$. This has been seen as a possible social distancing or distress around others coming through in writing. In the second piece of research, personal narratives recalling the day of the attack were collected alongside scores on the Impact of Events Scale – Revised (IES-R) (Weiss & Marmar, 1996) in the week following the events. IES-R scores were collected again five months later. The results suggested that lower use of first person plural pronouns was associated with higher PTSD symptoms in the immediate aftermath and higher levels of religious language were associated with higher PTSD symptoms after 5 months. Higher levels of anxiety language were also associated with lower levels of PTSD symptoms after 5 months. There are multiple potential explanations for these effects that could be suggested and that put these language measures forward as illustrations of emotional processes that occur after such an event. However, this piece of work looked at language use in 40 individuals and considered a minimum of eight language features at two different time points, essentially testing at least 16 associations when including the language features alone. It runs a risk of random findings and it is unclear at this point what the application of such research would be other than in illustrating how individuals express themselves following news of a traumatic event.

#### 1.3.1.2.4 Language measures and psychological scales.

A third focus of research into linguistic characteristics in mental health has been on the relationship between language use (in this case, word frequency) and established mental health scales and concepts. LIWC analysis of diary entries from over 4000 participants showed significant correlations of negative and positive affect as measured by the Positive and Negative Affect Schedule (Watson & Clark, 1999) with, respectively, levels of negative emotion language and of positive emotion language (Tov et al., 2013).

In a similar vein, a 2014 research project looked at the language used by mothers of children who were carriers of sickle cell disease during a semi-structured interview about their experience of genetic testing on their child. It was found that anxiety language use was significantly correlated with self-reported state anxiety, a measure taken prior to the interview, suggesting that elevated anxiety levels can be automatically picked up in language use (Ahmad & Farrell, 2014).

Lee & Cohn (2009) looked at correlations between language use and coping styles and concluded that more negative language when describing a stressful event was related to low problem-focused coping scores, while more insight language was associated with lower emotion-focused coping scores (Lee & Cohn, 2009). Insight language refers to words that express an individual's thought process and self-awareness. These are words such as 'believe', 'think', 'feel'. The idea is that individuals who use these terms more frequently have a greater understanding (or insight) into their own thoughts, feelings and interactions with the world around them. As with other LIWC categories it comes down to a frequency count of terms within the category that is then expressed as a percentage of all terms in a document.

Measures of problem-focused coping were based on the extent to which an individual searched for solutions to stressful problems whereas emotion-focused coping put the emphasis on an individual's management of their emotional response to a stressful situation as opposed to looking to affect the cause of the stress. Both effects above were found to be significant when considering the correlation between the frequency of the language category (negative language and insight language) and the score on both measures of coping style. However, despite being statistically significant the correlations estimated for these two effects were low, estimated at -0.14 for the association between negative language use and problem-focused coping and -0.19 for the association between insight language and emotion-focused coping scores. These results suggest that though there is a measurable association between language use and coping style but it does not seem that a difference in one will necessarily have a large impact on the value of the other and that going by language measures alone is unlikely to provide strong indication of coping style.

Two research groups looked at the language used by female victims of trauma or abuse. Holmes et al., (2007) studied the relationship between scores on LIWC categories and scale-based measures of pain and depression both at the beginning of and four months following the end of expressive writing therapy. The researchers hypothesized that an increase in the use of causal and insight language (within the cognitive mechanisms category of the LIWC) would be associated with improvements in pain and depression outcomes over the course of the writing sessions. They also hypothesized that higher levels of positive language would be associated with lower pain and depression scores and higher levels of negative language would be associated with higher pain and depression scores. Their results were not very conclusive, with measures of cognitive and emotion (positive and negative) words not being significantly associated with measures of depression. The level of pain reported by the participants was, however, negatively correlated with both negative and positive emotional language (Holmes et al., 2007) suggesting that higher emotional expression, whether positive or negative was associated with lower levels of pain. This can be seen to put forward the idea of expressive writing, and emotional expression, as a method of alleviating physical pain or the perception of physical pain.

The second piece of work involved the recording and analysis of 28 trauma narratives during a course of exposure therapy alongside a number of self-report scales such as the Beck Depression and Anxiety Inventories and the Quality of Life Self-Report scale. These narratives were split into three sections: pre-threat (up until the first expression that the person was in danger), threat, and post-threat (from the first expression of the realization that the danger had passed). Results suggested that in the pre-threat section, only the number of words relating to death or dying was significantly correlated to post-treatment psychopathology. In the threat section, the level of cognitive processing language (suggesting insightful or causal thinking: ‘cause’, ‘think’, ‘should’, ‘maybe’ etc) was negatively correlated with anxiety as measured by the Beck Anxiety Inventory and post-threat, positive language was negatively correlated with levels of anger reported within the Anger Expression Scale (AEX) (Alvarez-Conrad, Zoellner, & Foa, 2001). However, in this piece of work, seven LIWC categories were tested for potential associations with eight clinical measures and the sample size was only 28. This put it at risk of random findings as appears to be the case in a number of LIWC studies. There appears to be some inconsistency within similar populations. The results relating to cognitive language use or the significance of negative language are variable in the examples described here. Conditions for writing are often inconsistent both within and between studies and also with respect to topic, physical conditions and length of writing time. These elements will set the frame within which language is produced and used. Alongside the chance of random findings, the varying conditions of language production may go some way in explaining the inconsistent results. These in turn make it difficult to generalize across populations and writing contexts and limit the application of these results.

As can be seen from the variety of research work described above, the LIWC has been applied to mental health research in a variety of different ways. These include detecting differences in language use between groups with or without various mental health diagnoses and within group differences over times. A long-term clinical goal of this type of work may be in use of computerised language analysis as a diagnostic tool but at this stage, the research results appear unreliable and inconsistent. When looking at changes in language use over time, the LIWC has been put forward as a tool for monitoring progress in treatment. This may be especially useful within research as having a measure of treatment progress based on natural language may reduce the reliance on self-report scales. Within clinical application, this type of measurement of language variables may allow the identification of individuals who are not following an expected progression during their treatment and require extra support. Finally, a number of studies have looked to support the application of the LIWC in this way by measuring the association between standardized measures of psychological dimensions and a number of LIWC variables. Across each of the applications of the LIWC described here, a variety of research has been carried out but the work is very scattered in terms of methodology and the results do not come together as a coherent and consistent body of evidence at present. Results are often interesting and spark speculation over the processes behind significant associations but more rigorous work would help provide a clearer picture of how it can be applied.

### 1.3.2 Computer Assisted Language Analysis System

A different approach to linguistic analysis was followed in the development of the Computer Assisted Language Analysis System (CALAS) by Rush et al. in 1974. The system is grounded within case grammar and looks to categorise words and phrases by their grammatical use and context. Case grammar refers to the analysis of the relationships between the function of words and the semantic roles they play. It focuses on the number of 'deep cases' required by the verbs in a sentence. These cases are roles such as 'object', 'agent' or 'location'. For example, the verb 'to go' requires an 'agent' and can take a 'location', e.g. 'I went home.' The CALAS system was applied to look at differences in linguistic style between patients and therapists and within therapists in the context of dynamically focused psychotherapy (T. Anderson, Bein, Pinnell, & Strupp, 1999). The Computer Assisted Language Analysis System (CALAS) picked up on different verbal styles from therapists in high and low affect segments of text. High and low affect segments of text were identified by automatically counting the frequency of affective adjectives in the text using the Lexical Analysis of Verbalized Affect (LAVA), a computer program developed in 1995 based on an extensive affective language lexicon. In transcripts from patients who were successful in their treatment, stative verbs, verbs that describe a state rather than an action, were used more during high affect exchanges and action verbs were more present in low affect exchanges. It was also established that therapists spoke more efficiently than patients, that is to say conveyed more information in fewer embedded clauses (T. Anderson et al., 1999). This analysis method relies on recognition of grammatical features (block length and embedded clauses) and verb types and roles (stative, action, processing, experiencer affective or experiencer cognitive). The application of this system of linguistic analysis does not appear to have been followed up within mental health research, though the relevance and use of syntactical features can be seen across other methods of linguistic analysis. Further details on this piece of research will be included in the following chapter.

### 1.3.3 Content Analysis

#### 1.3.3.1 Psychiatric Content Analysis and Diagnosis

A second method of computerised linguistic analysis was a computerised version of the Content Analysis Scale, originally developed as a manual analysis method in 1969 by Gottschalk & Gleser. This method of linguistic analysis was developed as a set of coding or scoring rules to be applied to a five-minute speech sample elicited with vague instructions to talk about an important or dramatic event. It aimed to provide a measure of a number of psychological aspects such as anxiety, inward and outward hostility or depression, and was developed with the aim of being a tool to support diagnosis. Each scale was subdivided into themes (e.g hopelessness or self-accusation within the depression scale) and within each of these were defined a number of ways each of the themes could be expressed verbally. Within the hopelessness theme, for example, were the instructions to code ‘references to not being, not wanting to be, or not seeking to be the recipient of good fortune, good luck, God’s favor, or blessing’ and ‘references to self or others not getting or receiving help, advice, support, sustenance, confidence, esteem (a) from others, (b) from self’ as examples of hopelessness. Each type of verbal reference to a theme such as those provided above was given an associated weight (1 for both of the examples given). A score was then obtained by multiplying the frequency of each type of verbal expression by its associated weight. This was then divided by the total words in the sample and multiplied by 100 to provided a percentage score (Gottschalk & Hoigaard-Martin, 1985).

The primary limitation of this method was the requirement for manual coding and high inter and intra-rater reliability in order to be clinically applicable. The method was gradually computerised and the computerised version has been named the Psychiatric Content Analysis and Diagnosis (PCAD) (Gottschalk, Stein, & Shapiro, 1997). It relies on a dictionary and a set of scoring rules similar to those described above. However, it is very difficult to obtain clear information about exactly how the scales were computerized and scored. This is a primary criticism of the PCAD as it hinders appropriate evaluation and discussion of results obtained (Bantum & Owen, 2009).

A 1982 paper by Gottschalk and Bechtel reported specifically on the performance of the computerized anxiety scale within the PCAD as compared to scores from human raters based on verbal samples from 25 individuals. The results suggested that the PCAD (computerized) scores were consistently lower than human scores. The computerized scores on the anxiety subscales were, however, highly correlated, with an overall correlation of 0.85 for the six subscales (Gottschalk & Bechtel, 1982). This was a result that the authors put forward as supporting the validity of the computerized anxiety scale. Though this could be seen to support the relative performance of the Psychiatric Content Analysis and Diagnosis (PCAD) across anxiety subscales it suggests that it was not performing as intended in terms of absolute scores. This was followed by various projects reworking and improving the precision and performance of the PCAD and extending the number of scales the software could score. The scales included in the currently available PCAD manual are: Anxiety, Inward and Outward Hostility, Social Alienation-Personal Disorganisation, Cognitive and Intellectual Impairment, Hope, Depression, Human Relations, Achievement

Strivings, Dependency Strivings, Health-Sickness, and Quality of Life (*PCAD Manual*, 2016). Some scales, such as the anxiety scale, have been successfully translated into German (Berth, 2001).

A 2009 paper compared the PCAD with the LIWC and manual linguistic analysis of emotion in text and found that both computerised methods seemed to over identify emotion but that the LIWC performed better than the PCAD. This project recorded an average sensitivity (the number of identified emotion terms over the total number of emotion terms) of 0.88 for LIWC and 0.83 for the PCAD and a specificity (the number of terms identified that were correctly identified as non-emotion terms) of 0.97-0.99 for LIWC and 0.58 for PCAD. PCAD achieved a higher specificity (0.74) for negative emotional expression than was achieved for the other emotion categories. Despite this comparison with the LIWC, the paper supports the use of the PCAD to analyse textual data while highlighting the limitations to be aware of. This seems to be quite a lenient judgment of a coding method that was originally developed for diagnostic use. In addition to the obscurity of the coding mechanisms involved, the low performance as compared to another available method of computerized linguistic analysis suggests it requires further development before wide-scale application.

Nonetheless, the PCAD has been applied in a number of research studies, some examples of which will be provided here. The computerised content analysis scales have been tested as a potential part of the diagnostic process in a psychiatric outpatient clinic (Gottschalk et al., 1997). Scores on the scales applied (Anxiety, Hostility, Depression, Social-alienation and Person Disorganisation, Cognitive Impairment, and Hope) to speech samples of 25 outpatient showed significant correlations with self-reported scores on the Minnesota Multiphasic Personality Inventory (MMPI-2), the Symptom Checklist (SCL-90) and the Shapiro Control Inventory (Gottschalk et al., 1997). Significant correlations between content analysis scores and self-reported scores ranged from -0.43 to 0.45 and were found to support expected associations such as between the outward hostility scores (PCAD analysis) and depression scores (self-reported). The PCAD has also been used in the analysis of suicidal behaviour in Israeli veterans and terror victims in a paper published by Galor & Hentschel in 2009. This work picked up on differences in language use between individuals who had attempted suicide, those with suicidal ideation and control participants. As previously, the scores were based on five-minute speech samples of individuals describing an influential life event. Significant differences were present on 13 scales and subscales across the three groups (Galor & Hentschel, 2009). For example, scores on the hope scale were significantly lower for individuals who had shown suicidal behaviour as compared to a control group and there was a higher mean score on the total depression scale for individuals who had attempted suicide when compared to individuals who had displayed suicidal ideation. The application of the PCAD was put forward by the authors as an important method to identify individuals who might be at risk of PTSD or attempting suicide.

Though a number of pieces of work have applied the Psychiatric Content Analysis and Diagnosis scale, the number is limited when compared to those employing the Linguistic Inquiry and Word Count as a linguistic analysis method. In addition to this, the unclear methods behind it and its limited success in measuring emotional language as compared to the LIWC made it a less appropriate option when considering methods for analysis of language to apply in this research project.

#### 1.3.3.2 Computerised Referential activity

The importance of making linguistic analysis tools specific to context is something that has been taken into account in the following set of work. The therapeutic process in psychodynamic therapy has been the focus of the development of computerised tools with which to investigate it, with referential activity being a primary focus for an indicator of change in this context. Referential activity, also referred to as the referential process, has been put forward as an important process for a patient in psychotherapy. It involves making the association between a subjective experience and language by attributing words to a non-verbal experience, essentially the process of using words to describe a situation so that it can then be imagined by another individual (‘Referential Activity (RA) - The Referential Process’, 2015). For example, if an individual is recounting a childhood memory with specific details about the physical environment around them at the time as well as their feelings, this allows a listener to create their own image of the events more accurately, suggesting high referential activity in the speaker. It was originally coded manually and a dictionary to enable the computerized measurement of referential activity was developed in 1999. This dictionary was developed by selecting the 181 most frequent terms in texts that had been manually scored for referential function and scored either very high or very low (Mergenthaler & Bucci, 1999). It was developed on a dataset of 1018 documents making up a total of 368,590 words and followed a seemingly rigorous process. The developed dictionary was also manually checked with the removal of domain specific words that would not transfer to other contexts and its performance was compared to human raters tested on two independent sets of data. Computerised and manual scores were correlated with a score of 0.5, a promising result that nonetheless leaves room for improvement if it is to mimic human judgment. This computerized method of measurement of referential activity was then used in research in the area.

One application of this analysis looked at levels of referential activity in Thematic Apperception Test responses and the association of these with clinical outcomes and personality types in a population of psychiatric inpatients with a range of diagnoses including personality disorders, psychosis and depression (Fertuck, Bucci, Blatt, & Ford, 2004). The results suggested a different association between referential activity levels and clinical outcomes in the two personality types considered. These were anaclictic (emphasis on relatedness and empathy) and introjective (emphasis on self-control, self-worth and self-definition) personality configurations. They found that within individuals with an introjective personality configuration, increases in referential activity were associated with improvements in thought disorder outcomes, whereas the opposite was true in an anaclictic personality configuration. The last result was a surprise to the researchers and puts forward the idea that within one therapeutic context, opposite associations between language measures and clinical outcomes can be measured depending on a third factor. This highlights the influence of individual differences in personality that can affect the relationship between an individuals mental health and how this is expressed in the language they use.

#### 1.3.3.2.1 Computerised Reflective Function

A second area of research that aimed to automatically pick up therapeutic processes in language is work on reflective function. Reflective function differs from referential activity in that it involves the patient's ability to mentalise or put into words their and others' internal worlds, that is to say their emotions, motivations and beliefs. Reflective function is one process within Mergenthaler's Therapeutic Cycles model (Mergenthaler, 1997). This model posits that therapeutic progress is achieved through a specific sequence of states in the patient in therapy: starting with relaxing, where the tone is low on emotion and abstraction, this is followed by an experiencing phase, where the patient is high on emotional arousal, then into a connecting phase, where the patient is high on both emotion and abstraction and finally comes a reflecting phase in which the patient is reflecting on these feelings. Work to develop an automatic method to measure the stages of this cycle developed two dictionaries, one for emotional words and one for abstraction language (Lo Verde, Sarracino, & Vigorelli, 2012). The software was applied to transcripts from 206 sessions from 10 inpatients following psychodynamic therapy. The findings suggest that the connecting phase of the cycle is particularly important to progress in psychodynamic therapy (Bergmann, Villmann, & Gumz, 2008). This work provides interesting insight into the processes at work during therapy but with its small sample size, the findings need further support.

Further work on automating this reflective function measure was carried out with a paper detailing the development and assessment of criterion validity of a computerised measure of reflective functioning published in 2012 (Fertuck, Mergenthaler, Target, Levy, & Clarkin, 2012). High and low reflective function dictionaries were developed in a similar way as the referential function dictionary described previously. The frequent words and phrases in samples displaying either high or low reflective function were selected. The high Computer Reflective Function (CRF) measure was correlated with manual coding of Reflective Function (not based on language criteria) with a score 0.57 but the low CRF dictionary was less successful, providing little additional predictive power to Reflective Function scores based on high CRF. This work was carried out with data from 113 participants across two groups; a non-clinical control group and a group of individuals diagnosed with borderline personality disorder. The dictionary was developed on a sample of 18 texts and tested on a sample of 95 patient texts; this is a much smaller sample than was applied in the development of the referential activity dictionary. Though the results seem in line with those achieved for referential function in terms of agreement with human raters, the sample used seems too small to provide reliable results.

### 1.3.4 Machine learning – corpus-driven analysis

Much of the research surrounding automatic linguistic analysis in a mental health context has focused on developing algorithms for text or topic classification or other labels that could be automatically assigned to a section of text. These methods essentially involve the development of a computer program that can be run on a document or selection of documents and will provide a given output, most commonly a binary classification. Under the general term machine learning, this type of research relies on a large dataset as it requires a training set, a substantial development set and a testing set. These classifiers can involve one or a combination of rule-based classifiers or supervised and unsupervised machine learning algorithms. This area of research is rapidly developing and a variety of methods can be applied to one dataset, the primary obstacle in the area is access to the necessary quantities of data.

Text categorisation algorithms have been applied over a number of areas in mental health and across a variety of platforms ranging from clinical text to language used on twitter. Social media platforms are popular in this area of research as the computational linguistics community has a tendency to prefer working with data that is more openly available than confidential clinical data. However, some work has been done on more clinically focused topics as well as within therapeutic data. The following paragraphs aim to detail the variety of work that has been carried out in this field. The first set of work described involves research that has aimed to identify risks of mental ill health and evidence of symptoms of a variety of mental health conditions. These pieces of work primarily look at texts from online social media. The second part of this section provides some examples of work that has focused more on language within therapy sessions and how machine learning methods have been applied to these.

#### 1.3.4.1 Identifying evidence of Mental Health Disorders

##### 1.3.4.1.1 Depression

Picking up elements of language that might suggest depression or depressive symptoms in the author is a task that has received considerable attention in recent research, particularly with the popularity of social media and online communities that provide both an outlet for individuals suffering with depression and a source of natural language data for researchers. Neuman, Cohen, Assaf & Kedman (2012) looked to classify online texts as related to depression or not. In order to do this they developed a 'depression lexicon' that includes a range of phrases and words that individuals might use to describe a depressive state of mind without necessarily using the term 'depression'. This was combined with LIWC measures to create a classification tool that reached a correct classification level of 84%. The correct classification was manually determined by the authors from reading the texts (Neuman, Cohen, Assaf, & Kedma, 2012). Another two pieces of work were carried out on corpora extracted from depression focused online communities. The first looked at 400 posts within a depression community and judged them to contain high or low affect (positive of negative) based on the presence of emotional terms from the Affective Norms for English Words (ANEW). A classification tool was then built based on features from the LIWC and machine learning topics, and achieved, respectively, 78% and 60% accuracy in classification (Dao, Nguyen, Phung, & Venkatesh, 2014). Accuracy refers to the overall percentage of correct classifications. It is important to bear in mind that 'correct' classification here was determined by the Affective Norms for English Words dictionary that works on similar principles to the LIWC as each term within the 'affective' category is attributed a label of positive or negative. We might therefore expect them to be reasonably consistent.

The second piece of work was carried out on a larger sample from the same source with 38,401 posts from online depression communities as the clinical sample and 229,563 posts from other communities as the control sample. This piece of work used the aforementioned Affective Norms for English Words dictionary, LIWC, mood labels attached to the posts by the writer and machine learning topics as features in their classification tool. Topics are clusters of co-occurring words within a given data set. They found that including LIWC features alone achieved 88% accuracy in classifying the posts as depression focused or not and with the inclusion of the topics in addition to the LIWC category information, accuracy reached 93% (Nguyen, Phung, Dao, Venkatesh, & Berk, 2014). Accuracy refers to the overall rate at which the classification tool was correct. This type of work looks to find evidence of differences in language use between two sets of textual data. It is often developed with the aim of applying what was learnt to new datasets, for example that aren't clearly defined as depression-focused or not, in order to work as preventive measures in identifying individuals at risk. However, the knowledge that the online space an individual is writing in is a space for people with depression to express themselves is likely to mean that the language used is very different to that used in a more common online space. It is therefore uncertain how these results would transfer to a less clearly defined online environment and they require validation.

This form of linguistic analysis has also been used for more specific functions and recognizing individual symptoms such as classifying text that expresses automatic (dysfunctional) thoughts (Wiemer-Hastings, Janit, Wiemer-Hastings, Cromer, & Kinser, 2004). A specific tool was developed using a set of 149 labelled automatic thoughts. These were examples extracted from journal articles, handbooks and training manuals in cognitive therapy. Phrases such as ‘I will never be good at this’ or ‘this kind of thing always happens to me’ are examples of these automatic thoughts. Language features such as keywords based on content (e.g. expressions of failure), keywords based on grammatical parts of speech (e.g. adverb, emotional verb), syntax information and markers of tense were included in the development of a classification system. It was tested on a new set of 112 texts for which it correctly identified 77% of dysfunctional thoughts (Wiemer-Hastings et al., 2004). This result was seen as encouraging as the model performed at the same level with this new set as it had when the classifier was tested internally on the development data set. However, 77% still leaves large room for improvement and further validation in broader datasets if this tool is to be implemented in a clinical capacity. A related piece of research was carried out by Yu and colleagues where machine learning methods were applied to develop a tool that is able to classify the nature of negative life events (eg. Home, work, social, etc.) (Yu, Chan, Lin, & Lin, 2011). The focus in this paper was technical but its applications can be seen in the identification of types of negative life events or combinations of these that most affect the mental health of an individual.

These two pieces of work have focused on particular aspects that relate to an individual’s mental health. Automatic thoughts can be considered as manifestations of cognitive style or symptoms in language and negative life events are risk factors. Further work on identifying and extracting these features could be very useful in research in to the diagnosis, treatment and management of mental health disorders. For example, reliable identification of risk factors could support risk prediction work and identification of symptoms could support individuals and clinicians in the selection of appropriate treatment.

#### 1.3.4.1.2 Suicide and self-harm

One research area that has a history of being challenging in terms of risk prediction and prevention is suicide and self-harm. The larger data sets that have and are becoming more available and the capacity to work with large datasets that computational analysis provides means that there is great interest in the application of these methods within suicide and self-harm research. In the case of analysis of textual data two pieces of work led by Pestian had until recently been the primary areas where linguistic analysis had been applied. The first of these looked at the language contained in real and elicited suicide notes in order to train computer software to determine the differences between them (Pestian, Nasrallah, Matykiewicz, Bennett, & Leenaars, 2010). Elicited suicide notes were written by age, race and gender matched healthy controls, who were asked to write as if they were about to commit suicide. The subsequently developed machine learning algorithm was able to determine whether a suicide note was genuine or elicited in 78% of cases.

The same research group also looked into sentiment analysis within suicide notes in the context of the Informatics for Integrating Biology and the Bedside (I2b2) challenge, which, despite lacking evidence for clinical application is important to mention. It is a competition or challenge that takes place every year on a different topic. The 2011 challenge revolved around sentiment analysis of suicide notes. The challenge has given rise to a number of publications as individual teams each developed their method of text classification for the analysis of emotion in the text. Pestian and colleagues (2012) provided an overview of the results of this challenge, which was based on a corpus of 1319 suicide notes. Each text was manually annotated by 3 different volunteers who had been trained for the task. They were asked to label the text with 16 emotional labels including abuse, anger, blame, guilt, hopelessness, sorrow, happiness, love etc. This provided the gold-standard against which to measure the performance of each classification system. A variety of natural language processing systems were proposed. The most successful achieved a precision rate of 0.58 and a recall rate of 0.65 (Pestian et al., 2012). These results suggest reasonably low levels of success for the task provided and are reminders of the complexity and difficulty of analysis of emotions in written language.

#### 1.3.4.1.3 Post Traumatic Stress Disorder

Post-traumatic stress disorder is another area that has seen some applications of computational linguistics, with the focus being on identification of risk factors and diagnosis. The first to be considered is a piece of work on self-narratives that was published in 2012 by He and colleagues. The aim was to differentiate between individuals who had been diagnosed as with or without post-traumatic stress disorder (He, Veldkamp, & de Vries, 2012). 300 self-narratives in this case were collected as part of an online health survey and the participant was asked to describe the traumatic event they experienced and their symptoms. Inclusion criteria in the study were a diagnosis of having or not having PTSD from at least two psychiatrists and the experiencing of a traumatic event. The classification tool in this case was developed using a set of suggested keywords that were likely to be associated with PTSD narratives as well as labeling the phrases as PTSD or non-PTSD to allow the computer to learn which words would discriminate between the two classes of text. When the classification model was set to use a list of 25 keywords, it achieved an accuracy level of only 60%. This increased to 80% when the number of keywords was increased to approximately 50. No significant gain was made by further increasing this number. Though 80% accuracy may be a reasonably strong performance for classification tool within the context of other work described here, it still leaves a substantial amount of error that would be too large for any form of clinical implementation, despite emerging from narratives that contain clear instructions for focus on the traumatic event and symptoms. This suggests that its application on less focused narratives, as would be more likely if applied in practice, would be less successful. Better results would potentially be achieved with a larger development dataset. Nonetheless, it provides a good example of the kind of work that is being carried out and shows clear potential.

A later example developed a classification tool based on a manually built lexicon that aimed to identify web blog posts that referred to physical and emotional elements of combat exposure in members of the armed forces, differentiating these from control blog posts also written by members of the armed forces but without evidence of combat exposure (Konovalov, Scotch, Post, & Brandt, 2010). In this piece of work, the developed classification tool was determined to perform with a recall of 0.75 and a precision of 0.9. Recall refers to the proportion of posts that were identified compared with that should have been identified and precision refers to the proportion of posts within those that were identified that were identified correctly. Though the recall result leaves room for improvement, the precision rate here is promising. The work by Konovalov et al. (2010) focused on identifying combat exposure, a risk factor for PTSD. If further work concentrates on determining how the language used to describe personal experiences reflects the mental health of the writer, perhaps working in combination with methods applied by He et al., (2012), this could lead toward the development of a tool that aims to identify individuals who are at risk of developing post-traumatic stress disorder based on their verbal expression.

#### 1.3.4.1.4 Twitter-based diagnoses

More recent work has aimed to provide diagnostic labels for individuals based on the language they use on twitter. A selection of tweets was initially assessed for self-disclosure of schizophrenia. This process was carried out by automatically searching for phrases that include a variation of the character string ‘schizo-‘. These were then verified manually to confirm whether or not they were disclosures of diagnosis. This statement of a diagnosis by the twitter user was put forward as the correct classification that a classification tool to be developed would need to reach. 174 individuals declaring a diagnosis of schizophrenia were selected through this method and the dataset was matched with the same number of controls. A set of up to 3200 historic tweets from these users was then extracted. As has been the case in previous work, a combination of methods of linguistic analysis were applied to the data set including topic modeling (where clusters of words that appear in similar contexts are extracted), LIWC analysis and language clustering methods in order to develop a classification tool that achieved 82.3% accuracy (Mitchell et al., 2015). A similar piece of work by Coppersmith and colleagues carried out on data from the same set looked to distinguish a range of mental health difficulties including ADHD, Anxiety disorders, Depression, and Seasonal Affective Disorder (SAD). The best tool achieved 85% correct classification for anxiety detection with 10% false positives. This means that 85% of users were classified in the correct group (Anxiety disorder or no Anxiety Disorder) and 10% of the control users were wrongly classified as having an anxiety disorder. In the case of SAD, success was much lower with the correct classification rate reaching only 52% for a 5% rate of false positives (Coppersmith, Dredze, Harman, & Hollingshead, 2015). Though the results for the detection of anxiety are promising from a research perspective, the SAD results were very poor.

A final piece of work carried out within this data format involved a variety of the methods mentioned above as well as a more sophisticated method; supervised topic modeling. Supervised topic modeling means that documents are analysed accompanied by a label that can guide the topic modeling process. This label can take the form of a theme, a questionnaire score or specific diagnostic label, for example. Unsupervised topic modelling identifies clusters of terms based on statistical occurrence alone, but the supervised model has the additional information of the label to guide the modeling process and for which associated language is extracted. The topics can then be developed to be representative of those labels. The aim here was to classify depression-related language. The model was developed and trained on a set of expressive writing self-narratives with associated neuroticism scores. This was then tested using the Twitter dataset previously described. The classification tool emerging from the work achieved 75% recall (75% of users self-identifying with a depression diagnosis were identified) with one false positive for every 3 correct predictions; approximately 25% false positives (Resnik et al., 2015). The use of social media data within mental health research in this way is superficially attractive as it is convenient, often with easy and open access, is abundant, and often illustrative of the language that is used in everyday life. However, it is also a very noisy form of data to use as the context individuals are writing in can vary so greatly and it is difficult to determine what this context is. Furthermore, in the case of the three research projects mentioned above, there is no verification of diagnosis and classification is based on individual self-disclosure. This is likely to lead to a self-selecting sample of individuals who are willing to disclose and discuss their mental illness online, but also relies on the veracity of these statements. Given social stigma around mental illness, it is important to be aware that the language used by this sample is unlikely to be representative of a wider population of individuals with a mental health difficulty.

#### 1.3.4.2 Language in therapeutic data

Recent work by Imel, Atkins and Stevyers, shifts away slightly from a focus on diagnostic labels and symptom identification to looking into the characteristics of mental health treatment and how these differ between different forms of psychological therapy.  This type of research could also be seen to move towards identifying active components of treatment through the language used in therapy in order to further research these. The focus in the first example of this type of work was on patient-provider interactions in treatment sessions for a range of mental health conditions. Topic modelling was applied to a large corpus of text made up of over 1500 transcripts from a variety of treatment formats. These include cognitive behaviour therapy, psychodynamic therapy, drug management sessions and motivational interviewing. In topic modeling, documents are entered for analysis and clusters of co-occurring terms, called topics, are extracted automatically. These topics are then used as features in the development of classification models. A number of topics emerged within this piece of work including some around emotions, relationships and treatment with other topics focusing specifically on medication, pregnancy, appearance or conflict, for example. The themes for the word clusters were manually attributed based on the terms within them. The aim of the subsequently developed classification tool in this case was to discriminate between the different types of treatment (CBT, motivational interviewing, etc.). The results suggest that this was quite accurate with only 13.3% of the documents being misclassified (Imel, Steyvers, & Atkins, 2015). The model could be difficult to apply to a different dataset as it was tested on the same data it was developed on but the results are interesting nonetheless. Immediate application may not be obvious but this kind of work points to the idea that differences in treatment types can potentially be identified through the language used within these sessions. This is turn could be used as a measure of how closely a mental health professional is keeping to a prescribed treatment and subsequently whether adherence to this affects treatment success.

Three further pieces of work focused on a subset of the data set described above. These were transcripts from motivational interviewing for change in individuals with substance-abuse problems. Within motivational interviewing there is a coding method called the Motivational Interview Skills Code (MISC) that includes a number of therapeutic skills and behaviours. These are behaviours such as ‘affirming’, ‘questioning’, ‘reflection’, or ‘reframing’. A first piece of work in 2012 focused on the identification of ‘reflections’, when a therapist returns what a patient has said to them, sometimes rephrasing or adding to it. The classification tool developed in this case achieved an F-score (combined index of recall and precision) of 80%, suggesting a good result within the field of computational linguistics (Can, Georgiou, Atkins, & Narayanan, 2012). A later piece of work aimed to build on these results but aimed in this case to categorise patient language in transcripts from motivational interviewing as ‘change’ or ‘sustain’ talk. These are evidence in language of a patient looking to change a behaviour or resisting change of that behaviour (Tanana et al., 2015). However, the developed model performed below human reliability (Tanana et al., 2015), suggesting that the approach selected was not appropriate within this context. These are a few examples of the application of machine learning methods within therapeutic data. These will be covered in more detail in the next chapter along with further research that has specifically focused on language in therapeutic data.

#### 1.3.4.3 Electronic Health Records

The analysis of natural language notes on electronic health records has proved useful in classifying patients in a number of research projects. In one example, Perlis and colleagues compared the ability (to identify individuals with a diagnosis of major depressive disorder) of a classification tool that used billing data (diagnostic codes) with one that combined billing data with natural language notes included on the records. A selection of terms and phrases that might be indicative of major depressive disorder or absence of depression was devised by experienced clinicians. A classification tool was developed using logistic regression including diagnostic codes only at first and then including the natural language terms. The results suggested that the inclusion of these natural language terms in the classification model significantly improved the classification and identification of individuals with a major depressive disorder (Perlis et al., 2012). A second piece of work aimed to determine if a diagnosis of bipolar disorder could be identified using the natural language or free text sections of electronic health records (Castro et al., 2015). For this research records from 209 individuals were extracted and manually labeled by three mental health professionals with a label of either ‘bipolar disorder’, ‘no bipolar disorder’ or ‘not enough information’. At this point, the clinicians also had access to any diagnostic codes included in the record. As with the previous piece of work, the clinicians were requested to form a list of terms or phrases that would be indicative of the presence of absence of a bipolar disorder diagnosis. These expressions were then included as features in a classification tool developed using logistic regression. The tool was set up to reach 95% specificity (rate of true negatives), that is to say that of all the cases identified as not containing evidence of bipolar disorder, 95% of these should be correct. The result suggested that the classification tool reached an 85% positive predictive value. This is the proportion of true positives (bipolar disorder) over all those that were labeled as such. Given the high specificity, set to avoid false positives, this result seems quite strong. However, 132 of the 209 individuals included were diagnosed with bipolar disorder and this high prevalence may have inflated the strength of the classification tool as positive predictive value is sensitive to prevalence (Parikh, Mathai, Parikh, Chandra Sekhar, & Thomas, 2008). It is important to remember however that both the labels and diagnostic category were being provided by the same clinicians and that, in cases where a diagnostic code was provided, it is likely that the clinician writing the notes would have this diagnosis in mind and may be justifying it to some extent. This piece of work can be seen as looking at the consistency between the free text notes and diagnostic codes attributed rather than at the predictive value of the notes for a diagnosis. Furthermore, the language used in electronic health records is likely to be very different to that used by patients, suggesting that the same linguistic features used in modelling here may not apply to patient language. There is a variety of further work being carried out on the analysis of free text in electronic health records but these will not be covered in this thesis.

### 1.3.5 Conclusions and implications for research

The research area of linguistic analysis in mental health is both young and very diverse. A number of varying approaches have been applied, with the simpler, word count based, techniques being more popular with mental health academics. The very recent and growing area of machine learning methods still has a technical focus with few papers looking at the clinical applications of the complex algorithms they have devised. Combining all methods described here, there is great diversity of areas of mental health that have been studied. However, a majority of this research was carried out with relatively small sample sizes that cannot provide reliable or generalisable results without further replication and validation, and this is particularly true of the work carried out using the LIWC and computerised content analysis. The inconsistent results across linguistic analysis suggest that generalisability and external validity has been a concern in this field. Furthermore, taking population characteristics and context into account appears to be crucial to research design and interpretation.

One difficulty in working from results put forward by the literature review above is the range of data formats that have been considered. The majority of the work has been carried out looking at language use within personal narratives or similar document types. Given the role of context in shaping how we communicate, it seems important to consider how language has been analysed and what work has been done within the context of therapeutic dialogue. This is a question that Chapter 2 aims to answer with a systematic review of linguistic analysis within the context of treatment in mental health. Nevertheless, there are some obvious trends in the types of linguistic features that have been considered for research into this field. Most notable is the preference for the LIWC and within this, the analysis of affective language, pronoun use and the presence of social language (referring to friends, family or social actions). A number of other linguistic categories recur in previous research such as cognitive processing subcategories (insight, causality, certainty) and the use of negations. These features of language will therefore be considered in this research project, with the aim of exploring both how applicable the LIWC categories are to therapeutic dialogue data and how they relate to therapy outcome scores. Alternative measures of affective language will also be considered. The work on referential function suggests that work has begun looking at the therapeutic process. Though this was within the context of psychoanalysis, the idea of considering specific features of therapy or of the therapeutic process is one that is carried forward in this project.

This project carves out a specific area within this large field of linguistic analysis in mental health by focusing on analysis carried out using text mining and within transcripts from online cognitive behaviour therapy. The aims of the project are set out below.

## 1.4 Aims of the Thesis

Rationale: Throughout the literature review, results suggest that there are measurable features in language use that provide an indication of an individual's mental health status. Though the majority of these have been used to distinguish those living with and without a mental health condition (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Rude et al., 2004), language features have also been considered as indicators of progress over the course of treatment (Arntz et al., 2012). Given the growing popularity of computerised and text-based treatment options for mental health problems and the wealth of data provided by a service such as Ieso Digital Health, there is a lack of research into how such treatments work as well as an opportunity to look into this area to learn more about what is happening within treatment sessions. Beyond this, investigating specific language features as potential indicators of mental health status and treatment progress could provide a new method for monitoring and measuring this that would require no further input from a patient. Subsequent possibilities for adapting, changing and researching treatment are vast and the first step is to explore whether there are measurable features of language in CBT treatment that are associated with mental health outcomes. If specific features are found to be reliably associated with mental health outcomes, these could be considered as candidates for monitoring treatment progress, for effecting change or for adapting treatment to a patient's needs. The specific application will depend on the nature of the association between language and mental health outcomes, assuming there is one. As the nature of a potential association is unclear, this research project considers associations between language and mental health outcomes at different stages in the therapy process.

Firstly, an association between selected linguistic features during a therapy session and the most recently recorded mental health outcome is considered (outcome before session). This association primarily considers the linguistic features as potential markers of mental health status. Significant associations between measures of language use and recent mental health measures may suggest that these features are reflective of mental health state and could therefore be considered as possible candidates for progress monitoring. Secondly, an association between selected linguistic features during a therapy session and the closest future mental health outcome is considered (outcome before next session). This association involves an element of prediction and considers whether language use in a treatment session can predict short-term outcomes. This could provide a second opportunity for progress monitoring but also explores whether the presence of particular features may influence mental health outcomes. Finally, the third association considered is between language use early in treatment and mental health outcomes at the end of treatment. This association is concerned with longer-term prediction and could put forward early markers of treatment success as well as potential candidate features that influence outcome and therefore may suggest mechanisms for effecting change.

The overall aim of this thesis is to explore the potential of text mining in the analysis of online cognitive behaviour therapy and how it can be applied to learn about the therapeutic process within this context and improve service provision. This broad goal can be broken down into three elements:

1. To explore which linguistic features contained within online Cognitive Behaviour Therapy transcripts can potentially be measured with text mining methods and whether these are associated with mental health outcomes. The linguistic elements considered within this research project are based on three different sources:
    1.1. The Linguistic Inquiry and Word Count dictionary – exploring the application to this data set of a method of analysis that has previously

been used in mental health and how it can be adapted through text mining.

1.2. The Positive and Negative Affect Schedule (Watson & Clark, 1999) – adapting a different method of affect measurement to explore how it applies to this data and task.

1.3. The Cognitive Therapy Scale - Revised (Blackburn et al., 2001) – adapting four items within the scale in order to determine if adherence to the CBT structure could be quantified through text mining methods.

2. Based on the exploratory analysis described above, to develop predictive models of therapy outcome based on demographic, baseline and linguistic data and assess the contribution of the linguistic features extracted through text mining. Statistical analyses will be used to consider the following aims:

2.1. To establish whether language features can be considered markers of mental health status by investigating the association between language use in a treatment session and the mental health outcomes recorded before the session.

2.2. To establish whether language features can be considered as short term predictors of progress in online text-based CBT by investigating the association between language use in a treatment session and the mental health outcomes recorded before the next.

2.3. To establish whether language features early in treatment can be considered predictors of outcome at the end of treatment by investigating the association between language use in the first two treatment sessions and end of treatment outcome scores.

2.4. Additionally, associations between linguistic features and time to drop-out will be explored by investigating the association between language use in a session and drop-out after the session.

3. To understand how text mining methods can be applied to online CBT transcripts in order to assist future research into the treatment provided, such as the active ingredients of online CBT or elements that influence

change in a patient. This aim is primarily methodological and looks to support future research so as to continue to learn about and potentially improve the service provided.

This research project is primarily exploratory but it has potentially vast clinical implications. Evidence of associations between linguistic features during treatment and mental health outcomes may suggest either markers of treatment progress or factors that may be impacting those outcomes. Successful predictive models could allow patient cases that look to be at risk of poor outcomes to be brought to the attention of the service provider. No clear intervention for these is suggested at present with the focus being on determining whether such models can be developed.

# Chapter 2. Systematic review of the literature on computerised linguistic analysis in therapeutic dialogue data.

## 2.1 Introduction

The previous chapter provided an overview of a large body of work in which computerised language analysis has been applied within mental health research. It paints a varied picture with a number of designs, analysis methods and data formats being considered. This ranges from LIWC analysis of personal narratives in an eating disorders unit (Wolf et al., 2007) to algorithms looking to classify automatic thoughts as dysfunctional or not (Wiemer-Hastings et al., 2004). Research into language use in mental health also brings together the research fields of mental health and computational linguistics, which can follow differing norms, thus adding to the diverse picture.

Given the scope of the literature covered in Chapter 1 and the rapidly changing nature of this field of work, it is necessary to include a systematic review of relevant work. The primary restriction here will be on the type of data used for analysis. The reason for this is twofold. In the first instance, it became apparent from the literature reviewed in Chapter 1 that the context in which textual data is produced can strongly determine the appropriate methods and results and that this should be considered in any interpretation. This is not counter-intuitive as most individuals will adjust their language to the context in which it is being used (Hymes, 1967). Secondly, reviewing research on language used specifically within a therapeutic setting will provide a more detailed picture of the specific context within which the research completed in this thesis was conducted.

Though a large amount of the work reviewed here will be taken into account throughout the rest of this research project, the primary aim of the review was not to inform method as this was at least partially determined a priori.

This was due to the nature of the project as an industrial collaboration exploring the application of text mining within mental health transcript data. Furthermore, a large amount of the work reviewed here was carried out and published after the project had started. Therefore the focus was on exploring how different analytical methods have been applied to therapy data, in order to place them in context with the method used here. It follows the structure and methodology of a systematic review.

## 2.2 Aims

This review aims to answer the following question: How have computerised language analysis methods been applied to textual data emerging from a treatment session for a mental health condition? Thus, it aims to develop a detailed picture of the methodological approaches that have been applied in the analysis of language used in mental health treatment sessions. This is being done in order to situate this research within the current body of work in the field. This review will bring together work from the mental health field and that of computational linguistics.

## 2.3 Method

### 2.3.1 Eligibility criteria:

a) The research must focus on language used by individuals receiving or providing treatment for a mental health condition.

b) The language analysed must originate from a treatment session that involves a conversational exchange between a mental health professional and the individual seeking help.

c) Selected pieces of research must contain an element of computerised textual analysis.

d) Any research design was considered, including secondary analysis of data, with the exception of case-studies. Observational studies and secondary analyses of data were the most frequently expected data formats.

### 2.3.2 Information sources

Literature searches were carried out on PsycInfo, Web of Knowledge and PubMed for work published at any time since 1992. This date was chosen as this is when the paper by Garfield and colleagues reviewing the application of Natural Language Processing in Psychiatry was published ((Garfield et al., 1992) - described in Chapter 1). This paper provided a review of the applications of textual analysis methods within mental health up until that date.

It was expected that most papers would be published in English and translations were sought when this was not the case. Additionally, the anthology of the Association of Computational Linguistics, an archive of proceedings from Association of Computational Linguistics conferences, and Ethos were searched manually in order to find relevant work from the computational field and doctoral theses.

### 2.3.3 Search strategy

Keywords around the two main concepts of mental health and linguistic analysis were generated based on keywords associated with known relevant work and in consultation with an information scientist at UCL.

The final search terms were:

("linguistic analysis" or "computational linguistics" or "computer* analysis" or "text mining" or "machine learning” or “natural language processing” or “nlp”) AND (Anx* or depress* or panic or phobi* or agoraphobi* or stress or dysthimi* or psychosis or ocd or "obsessive compulsive" or schiz* or "affective disorders" or addiction or dependence or bipolar or exploded MeSH term: Mental Disorders)

### 2.3.4 Data management and synthesis

Relevant literature was downloaded and managed using Zotero (Stillman, Kornblith, & Cheslack-Postava, 2013) and the relevant information extracted into a Microsoft Office Excel spreadsheet.

The data contained in the relevant literature will be collated into a narrative synthesis as the focus is primarily on methodology. Research will be grouped by method and summarized and briefly discussed within these groups. A general discussion will then follow.

### 2.3.5 Results of literature search

The literature search was carried out in December 2015. The initial literature search returned over 5000 papers across the three databases with high numbers of irrelevant papers. This figure does not include papers emerging from the manual search of the Association of Computational Linguistics (ACL) anthology, the archive of proceedings from ACL conferences. The search was then restricted by including the keyword ‘language’ as a required term. This search returned 1580 results across the three databases. 242 potentially relevant papers were selected from manual reviewing of titles, reducing to 186 after the removal of duplicates. Following further reviewing focusing on the therapeutic nature of the textual data included, this number was reduced to 39.  10 papers were added to the selection from the ACL anthology. 35 papers were excluded for not meeting the eligibility criteria after full reading of the papers. See Figure 2-1 for a diagram of paper selection.

The final number of papers included in this review is 14.

**Figure 2-1 Flow diagram of search results**

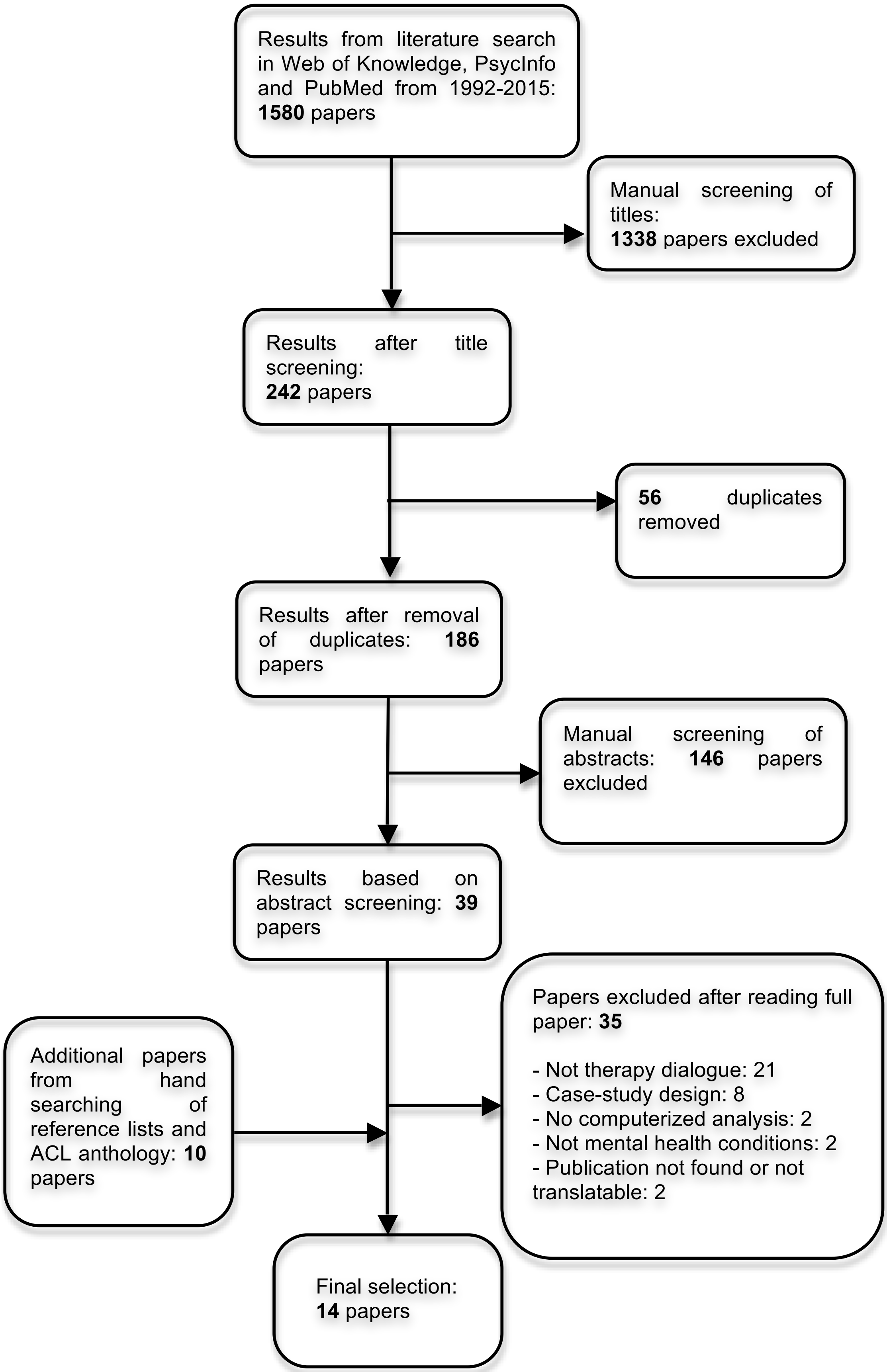

**Table 2-1 Summary table of selected study characteristics**

| Citation | Journal | Origin of sample | Mental health disorder being treated (if applicable) | Therapy type and format | Linguistic Analysis method(s) |
|---|---|---|---|---|---|
| Anderson, T., Bein, E., Pinnell, B., & Strupp, H., (1999) | Psychotherapy Research | 32 psychotherapy session transcripts | Unclear | Face-to-face, brief, dynamically focused psychotherapy | 1) Lexical Analysis of Verbalized Affect 2) Computer-Assisted Language Analysis System (CALAS) |
| Atkins., D.C., Steyvers, M., Imel, Z.E., & Smyth, P. (2014) | Implementation Science | 148 sessions from 5 different previously conducted randomised controlled trials | Drug and alcohol abuse. | Motivational interviewing | Semi-supervised labelled topic modelling |
| Can, D., Georgiou, P. G., Atkins, D.C., & Narayanan, S.S., (2012) | Interspeech 2012, Conference | 57 sessions from 3 intervention studies | Drug and alcohol abuse. | Motivational Interviewing | Automatic extraction of n-grams |
| Fontao, M.I., & Mergenthaler, E., (2008) | Psychotherapy Research | 42 hours of group therapy recorded and transcribed | Eating disorders | Group psychodynamic therapy | The Therapeutic Cycles Model program |
| Haug, S., Strauss, B., Gallas, C., & Kordy, H. (2008) | Psychotherapy Research | Transcripts from 200 chat sessions | Mixed diagnoses | Aftercare internet chat group intervention after inpatient treatment. | Statistically-based measures of activity, indegree, outdegree, indegree therapist and LIWC |
| Howes, C., Purver, M., McCabe, R., Healey, P.G.T., Lavelle M., (2012) | Proceedings of SIGDIAL[1] 2012 | 131 recorded and transcribed outpatient consultations. | Schizophrenia and schizoaffective disorder | Outpatient consultation with psychiatrist | Automatic extraction of turn-level features (speaker, number of words, filler words etc.) and unigrams |
| Howes, C., Purver, M., McCabe, R., Healey, P.G.T., Lavelle M., (2012) | Proceedings of SemDial[2] 2012 | Unclear | Schizophrenia and schizoaffective disorder | Outpatient consultation with psychiatrist | Automatic extraction of turn-level features (number of words, filler words etc.) and unigrams |

| | | | | | |
|---|---|---|---|---|---|
| Howes, C., Purver, M., & McCabe, R., (2013) | Proceedings of IWCS[3] 2013 workshop | 138 recorded and transcribed outpatient consultations | Schizophrenia and schizoaffective disorder | Outpatient consultation with psychiatrist | Topic modelling |
| Howes, C., Purver, M., & McCabe, R., (2014) | Proceedings of CLPsych[4] 2014 | 882 transcripts from online CBT from 167 patients | Mild to moderate anxiety and depression | Text-based Cognitive Behavioural Therapy | Topic modelling, sentiment measures, and automatic extraction of high-level features and n-grams |
| Imel, Z.E., Steyvers, M., & Atkins, D. C., (2014) | Psychotherapy | 1553 psychotherapy and psychiatric medication management sessions | Variety of conditions | Varied: Motivational interview, psychodynamic, experiental/humanistic, CBT etc. | Parts of speech tagging and topic modelling |
| Tanana, Hallgren, Imel, Atkins, Smyth & Srikumar (2015) | Proceedings from CLPsych[4] 2015 | 356 sessions from 6 different studies | Drug and alcohol abuse. | Motivational interviewing | N-grams and word vectors pre-trained by the Glove Model (Pennington et al., 2014). |
| McCarthy, Mergenthaler & Grenyer (2014) | Psychotherapy Research | Transcribed audio recordings of 20 psychotherapy sessions | Personality disorder with depression | Psychodynamic therapy | The Therapeutic Cycles Model program |
| Xiao, Imel, Panayiotis, Atkins & Shrikanth (2015) | PLOS One | 190 sessions from a multi-site randomized controlled trials | Drug and alcohol abuse | Motivational interviewing | Words used to build models through Support Vector Machine modelling with LIBSVM toolkit |
| Van der Zanden et al., (2014) | Journal of Affective Disorders | Chat sessions from 234 patients in randomised controlled trial | Depression and anxiety. | Master Your Mood online CBT group therapy. | LIWC |

1. SIGDIAL: Special Interest Group of Discourse and Dialogue. 2.SemDial: Workshop on Semantics and Dialogue. 3. IWCS: International conference on computational semantics. 4. CLPsych: Workshop on Computational Linguistics and Clinical Psychology.

### 2.3.6 Summary of papers

## 2.4 Results

The fourteen papers selected for this focused review are diverse in population studied, therapy applied and analysis methods used. As shown in the previous chapter, the approach behind the work was on a spectrum that ranges from computational linguistics (with very technically advanced but less applied work) to mental health (with more applied but less technically sophisticated methods). The majority of the work involved collaboration between the two disciplines with varying levels of influence of each one, possibly dependent on the publication platform selected. The variation in the research work done in this area makes direct comparison a complex task. With the aim of providing a picture of how linguistic analysis methods have been applied in therapeutic dialogue, I will summarise and discuss the research work that has been carried out. The research will be broadly grouped into two groups: dictionary-based analysis methods and the development of classification models for mental health outcomes, generally relying on machine learning methods. Table 2-1 provides summary information about the source and design for each paper selected. A summary table of analysis and results can be found later in the chapter (Table 2-2).

### 2.4.1 Dictionary-based approaches

As was the case in the background chapter, the first section concerns research that applied dictionary or frequency-based methods of textual analysis. This means that the method applied involved the measurement (mostly frequency-based) of given linguistic features that were generally defined within a dictionary setting out categories or sets of words. The association of these with recorded or reported mental health outcomes was then analysed statistically.

#### 2.4.1.1 Computer-assisted Language analysis system by Anderson et al., (1999)

The first piece of work to be covered here stands apart from the other dictionary-based approaches as it used a different system with greater focus on grammatical roles and interrelationships as opposed to semantic categories organised by meaning. This work was briefly covered in the background chapter. The data were a set of transcripts from psychotherapy sessions for 32 patients who had taken part in the Vanderbilt II study, a study focused on the effects of therapist training in psychodynamic psychotherapy (Strupp, 1993). No specific diagnosis was provided for the patients included (T. Anderson et al., 1999).

Anderson et al., (1999) were primarily concerned with measures of grammatical features in high and low affect segments of the recorded therapy sessions. High and low affect segments were identified using the Lexical Analysis of Verbalized Affect (LAVA) programme that was developed by the authors and based on a taxonomy of 500 affective terms published by Ortony, Clore and Foss in 1987 (Clore, Ortony, & Foss, 1987). Each transcribed segment of data was split into thought units (a portion of speech expressing one complete thought) (Henry, Schacht, Strupp, 1986) and each thought unit was attributed a frequency score of affective terms. For each patient case, the segment of 25 consecutive thought units with the highest relative frequency of affective terms and the segment of 25 consecutive thought units with the lowest relative frequency of affective terms were selected as the high and low affective segments for that patient case.

Within these segments, the authors then obtained measures of verb usage and stylistic complexity with the Computer-Assisted Language Analysis System (CALAS), a system first suggested by Pepinsky in 1978 (Pepinsky, 1978). CALAS was able to identify noun and verb clauses and categorise verbs by whether they were stative verbs (describing a non-causal relation), action verbs (causal relationship with specification of an agent) or process verbs (causal without specification of an agent) (Anderson et al., 1999).

Stylistic complexity is a measure that combines block length (a measure of the amount of information in a text) and the number of embedded clauses, essentially measuring efficiency of communication by looking at the ratio of information to the number of clauses (Pepinsky, 1985). This analysis system was used to count the frequency of different verb types and levels of stylistic complexity in the high and low affect segments. Statistical analysis was then carried out using multivariate analyses of variance to look at the relationship between the therapy outcome (good or bad), the level of affect in segments of therapy transcripts (high or low), the speaker (patient or therapist) and the levels of verb usage and stylistic complexity.

The results suggested that in cases with poorer outcomes, therapists tended to use more stative cognitive verbs (e.g 'think', 'believe'), as opposed to stative affective verbs (e.g 'feel', 'desire') in high affect segments in comparison with language used by therapists in cases with a good outcome. It was also found that there were differences in speech patterns between therapists and patients with therapists using more stative (non-causal, descriptive) verbs in high affect sections and more action verbs in low affect sections as compared to the patients. From the measures of stylistic complexity it was also found that therapists appeared to be more efficient in their language use as compared to patients with more information being conveyed in fewer embedded clauses. Confounders such as differences in education or therapist familiarity with the subject matter may provide explanation for this result. Speech complexity did not appear to differ significantly between good and poor outcomes.

The primary finding in this piece of work therefore focuses on therapist use of cognitive or affective language when a patient is using high affect language. Worse outcomes occurred when a therapist responded to high patient affect with more cognitive language. This could be interpreted as distancing from the emotional tone and a therapist not being 'in the moment' with a patient. The authors note that these differences were measurable when they were within the bounds of high or low affect segments but would not have been noticeable if scores were averaged across the therapy session. This puts forward an argument for splitting data into sections to allow the study of language as an indicator of behaviour or response to particular situations. It also supports the suggestion that language is very context-specific and that any analysis tools developed should be context-specific. The method used in Anderson et al., (1999) was an original approach to linguistic analysis within mental health. However, it also appears that the method has not been replicated and that there are few comparable studies, perhaps suggesting difficulty in its application. Additionally, the small sample size of 32 and use of multivariate analyses suggests a necessity to replicate this work.

#### 2.4.1.2 Emotion-abstraction patterns and the Therapeutic cycles Model in Fontao & Mergenthaler (2008) and McCarthy, Mergenthaler, & Grenyer (2014)

Two pieces of research selected for this review applied the Therapeutic Cycles Model (Mergenthaler, 1996) to textual data emerging from psychological therapy sessions. The Therapeutic Cycles Model is a computerized tool based on the idea that key moments of change or progress in psychological therapy, specifically psychodynamic therapy, are brought about through a cycle of linguistic behaviour. In this patients move through stages of expressing themselves with different levels of abstract and emotional language to make connections and create new understanding of their experiences. Further details of the model can be found in Chapter 1 (1.3.3). The computerized analysis relies on a dictionary through which words used are categorised as emotional, abstract or neither.

The first of the two studies applying this method worked with transcribed group therapy sessions from a group of female patients diagnosed with an eating disorder (Fontao & Mergenthaler, 2008). They were following a psychodynamically-focused course of therapy and the group was made up of a core group of five patients, though some sessions were attended by up to 8 patients. Forty-two hours of recorded therapy were transcribed and used in the study, equating to forty-two sessions. The Therapeutic Cycles Models software was applied to measure emotion and abstraction patterns and identify stages of the cycle in the data. Fontao & Mergenthaler (2008) then conducted analyses of variance in order to determine any association between these and therapeutic factor ratings that were coded manually following the Kiel Group psychotherapy process scale (Rohweder & Wienands, 1993). They confirmed their hypothesis that language patterns changed with the therapeutic processes. Specifically, they found that the manually coded therapeutic process *insight* was associated with the automatically coded pattern of *connecting* (high emotion and high abstraction) and that the therapeutic process *catharsis* was associated with the automatically coded pattern *experiencing* (high emotional tone, low abstraction) (Fontao & Mergenthaler, 2008).

McCarthy et al., (2014) applied the same method of linguistic analysis to a set of transcripts from 20 patients completing a course of psychodynamic therapy for depression in individuals with a personality disorder. Patients were selected for inclusion based on their 12-month follow-up outcome scores with 10 individuals selected as a poor outcome group and 10 selected as a good outcome group (McCarthy et al., 2014). In each case the third therapy session was transcribed and included in the analysis. Each sixty-minute session was split into three twenty-minute parts (beginning, middle and end). Similarly to Fontao & Mergenthaler (2008), McCarthy et al. (2014) also used analyses of variance but this time for between group comparisons of language use from individuals with good or poor therapy outcomes. Their results suggested that more improved patients spent significantly more time *connecting* (high emotion and high abstraction) in the first two sections of a therapy session and significantly more time *relaxing* (low emotion low abstraction) in the final part of the session. Additionally, the least improved patients spent significantly more time *connecting* in the final session. These results suggest that the timing of the stages of the therapeutic cycles and allowing for sufficient time to relax after a connecting stage prior to ending a therapy session may be important to progress in psychodynamic therapy.

Both pieces of work described above put forward results that allow some insight into the process of change in psychodynamic therapy and the type of language patterns that may be associated with these therapeutic factors. The model on which these analyses are based is grounded within psychodynamic theory and it is uncertain how the results would generalize to other forms of psychological therapy. This raises questions for further research around adapting methods to alternative treatment formats but also limits the comparability with the research carried out within this thesis. Furthermore, both pieces of work described here have their own associated limitations. Fontao & Mergenthaler (2008) worked with group chat data meaning that in the context of this review, results are not easy to generalise to individual online therapy, but also face-to-face therapy. Furthermore, Fontao & Mergenthaler's work followed one specific therapy group throughout their course of therapy, bringing it close to a case-study type design despite the presence of a small group of individuals (Fontao & Mergenthaler, 2008). Sample size is a potential issue for both pieces of work: McCarthy et al. (2014) used a small sample of 20 patients, with 10 in each group, a small sample made-up of patients selected based on extreme good or poor outcomes. The selection method, small sample size and questions about how those falling between the two extreme outcomes would behave linguistically make these results difficult to rely on or draw general conclusions from.

#### 2.4.1.3 LIWC measures and group therapy settings

The final two pieces of work applying dictionary and frequency-based measures of language included the application of the previously described Linguistic Inquiry and Word Count (LIWC) dictionary on data from two online group therapy settings. The first was an online group aftercare program for individuals who had received psychological treatment for a combination of mental health issues including depression, anxiety, stress, and behavioural and personality issues (Haug, Strauss, Gallas, & Kordy, 2008). This dataset was comprised of 200 chat sessions from four different groups, including a total of 130 participants. Participation was rolling so that if an individual left a group, another could be included, meaning that participants were not necessarily constant throughout the research study. In this piece of work, Haug et al. (2008) defined new measures of group dynamics such as the amount of activity an individual had in a group chat (the number of contributions), *indegree* (a measure of how many times other group member referred to a given participant, and *outdegree* (the number of times a participant referred to other group members) as well as using the LIWC dictionary to measure 52 variables including first person pronouns and communication language. Correlation analyses looked at the associations between scores of group therapeutic factors or group relationship, that had been provided through manual scoring of transcripts, and the measured linguistic features (LIWC, 'in/outdegree' and frequency of activity). The results from the study suggested that increased use of first person singular pronouns was associated with a lower quality of group relationship. Consistent with this, the opposite association was found for first person plural pronouns, which was found to be associated with a higher quality of group relationship. In terms of association with symptom measures, higher *indegree* (other users referring to patient), *therapist indegree* (therapist referring to patient) and *activity* were associated with lower symptom severity (Haug et al., 2008).

The next piece of research on online group therapy data involved an online psychotherapy course called 'Master Your Mood' aiming to guide participants through a series of modules to improve their mental health with regular group chat sessions to monitor and discuss progress. Facilitators guided sessions on a group online chat. The data collected was made up of application forms from 234 participants and the chat transcripts from those within the original participant pool who completed the course (Van der Zanden et al., 2014) with the data of interest here being the chat transcripts. The linguistic features measured in this work were the number of words typed by each participant and categories from the Linguistic Inquiry and Word Count dictionary, described previously. Seven variables were selected for analysis: *First person singular pronouns, positive emotions, negative emotions, causation,* *insight, discrepancy* (‘would’, ‘should’, ‘could’, ‘conflict’, ‘wish’) and *social processes* (‘share’, ‘we’). Correlation and regression analyses were then carried out looking at the associations between the language features and the main outcome: *mood mastery*, a measure of an individual’s belief in their ability to control their environment, and therapeutic alliance and symptom severity scores (Van der Zanden et al., 2014). Higher mood mastery was put forward as an indicator of better mental health outcomes. During treatment an increase in discrepancy language was associated with a decrease in depression levels. A result associated with baseline levels of discrepancy language helps to contextualise this result. Higher use of discrepancy language on the application form was associated with higher *mastery* levels before treatment and fewer discrepancy words at baseline was associated with greater improvement during treatment. These results may be describing the same phenomenon that use of discrepancy language is associated with *mood mastery*. This was a surprising outcome as it was suggested that *discrepancy* language (‘should’, ‘would’, ‘could’) would be associated with worse mental health outcomes due to being an indication of a disjoint between an individual’s actual and desired circumstances. However, it was suggested that *discrepancy* language in this therapeutic setting was associated with future ambitions as opposed to current shortcomings (Van der Zanden et al., 2014). Beyond looking at mental health outcome scores and language during therapy sessions, this piece of work also considered adherence and attendance. Though not considering language during treatment, it is interesting to note that better attendance was positively associated with the number of words used by the patient on the application form (Van der Zanden et al., 2014). This echoes results from Haug et al., (2008).

Both pieces of work described above suggested associations between measures of language use and mental health or therapy outcomes. Additionally, Haug et al. (2008) put forward original measures of involvement in group therapy with their ‘indegree’ and ‘outdegree’ measures. One common result of the two studies is the suggestion that how much text a patient contributes, either at application or during treatment is positively associated with attendance or improved symptoms, putting greater involvement and activity forward as a possible contributor to treatment. However, the authors of both pieces of work note that their results require further investigation and replication. Van der Zanden et al. (2014) tested a high number of possible associations and raise the possibility that some associations were significant by chance (Van der Zanden et al., 2014). Furthermore, the analyses carried out provide little information about what the mechanisms behind the associations between language features and outcomes might be. Further work on this front would both strengthen results and provide a route for clinical application.

#### 2.4.1.4 Brief discussion of frequency based methods

Five papers were reviewed in the section above. All consider frequency-based measures of linguistic features and their association with either mental health treatment outcomes or therapeutic factors. The therapeutic factors were either coded manually from the reading of transcripts or recorded through questionnaires completed by the patient or clinician. Completing this type of research requires pre-defined features of language to be measured and tested, which can be a limitation in itself due to the subjectivity of the process. Furthermore, the discovery of new and useful linguistic features can be labour-intensive with an element of trial and error. This stands in contrast to some of the machine learning methods that will be covered in the next section, which allow greater scope for language features to emerge from the text without prior hypothesis. However, these pre-defined, often human generated, language features are generally more straightforward to interpret when evidence of significant associations is found. For example, the results in McCarthy et al., (2014) suggested that a period of *relaxing*, in which a patient was not using highly emotional or abstract language (after spending time making connections from their experiences, evidenced by both highly emotional and highly abstract language) was important to good therapy outcomes. This can be understood reasonably easily in context. Therefore, while pre-selection of linguistic features may be a lengthy and subjective process, it can lead to more interpretable and applicable results.

Two different approaches to the segmentation of the textual data were put forward within these five studies. The first was to consider all the text provided by an individual within a treatment session as one document for analysis, with score of language use measured across that session as a whole. The second involved breaking up each session by splitting it into potentially meaningful parts. McCarthy et al., (2014) split their data chronologically whereas Fontao & Mergenthaler (2008) focused on analysis of segments in which affect (emotional language) was determined as either high or low depending on the frequency of emotional terms within it. In the case of McCarthy et al. (2014), the authors suggested that, had the data not been split in this way, significant differences between transcripts from good and poor outcomes would not have been measured. These results therefore suggest that paying attention to when linguistic variables are used in a session may be an important part of this type of research.

A recurrent theme across the pieces of research presented above is a concern with small sample sizes or the carrying out of multiple analyses on the same dataset. Most of the work is presented as exploratory and there are only a few overlapping methods between the studies included. Even when the same linguistic analysis method is applied (Fontao & Mergenthaler, 2008; McCarthy et al., 2014), the data format and outcome measures vary greatly. This was to be expected in this field but is none the less an important point to take away as some of the work presented requires replication and further elucidation such as the positive association between the frequency of *discrepancy* terms in group therapy sessions and better mental health outcomes (Van der Zanden et al., 2014).

Overall, dictionary-based methods of labelling of words appear practical and, once a given dictionary is established and validated, can be considered objective and used as a tool across multiple data formats and research projects. In this aspect, they appear to be a useful research tool as they can allow the comparison and contrast of different data sources. However, these dictionary-based methods remain frequency counts of individual words. This means there are inevitable limitations, as subtleties of linguistic context cannot be taken into account and ambiguous words or homonyms are likely to be miscategorised. They can be useful as a research tool, provided that these limitations are taken into account and the interpretation of the results is made with clear awareness of what is being measured.

#### 2.4.1.5 Classification problems and machine learning methods

The papers that will be covered in this second section include a variety of machine learning methods as applied to linguistic analysis. The majority of this work emerges from the field of computational linguistics. Machine learning concerns pattern recognition and is applied in these studies in two ways. Firstly, in the extraction of linguistic features such as in topic modelling where clusters of words that co-occur in textual data are extracted as a 'topic' or a distribution of words over a document. These features can then be used, along with any other measurable features (e.g LIWC measures, individual word frequencies, time data), in the second application of machine learning, which is to develop the algorithm or prediction model. Research questions are often put forward as classification tasks, either binary or multiple classification, with the developed algorithm identifying whether a specific portion of text or document belongs to a class or not. For example, an algorithm developed to solve a negative sentiment classification problem will aim to identify whether or not a piece of textual data can be considered to contain language associated with negative sentiment. The models or algorithms developed to perform these operations are complex statistical processes, the details of which go beyond the scope of this thesis. Furthermore, the details of the algorithm are often inaccessible in publications and can be very difficult or impossible to interpret in a meaningful way depending on the method applied. For these reasons, though the structure of a model may be mentioned, it's mechanism and functioning will not be detailed.

### 2.4.1.6 Classification work with a focus on Motivational Interviewing

The first set of research studies that applied machine learning methods to solve a variety of classification problems is made up of five papers that report on work on overlapping data sets (Atkins, Steyvers, Imel, & Smyth, 2014; Can, Georgiou, Atkins, & Narayanan, 2012; Imel et al., 2015; Tanana et al., 2015; Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015). They comprised a selection or combination of transcripts from up to six Motivational Interviewing intervention trials for alcohol or drug abuse. Additionally, the work by Imel et al., (2015) included textual data from the general psychotherapy corpus maintained by the 'Alexander Street Press' (htto://alexanderstreeet.com) which brings together transcripts of psychotherapy and drug therapy from a range of theoretical backgrounds including sessions by Albert Ellis and Carl Rogers and others who were developers of different treatment approaches (Imel et al., 2015). Some of these were originally published as sample sessions and training materials.

Four of the pieces of research within this group aimed to predict Motivational Interviewing Skills Code (MISC) labels. The MISC is made up of 12 codes or labels that are used to manually assess therapist behaviour and patient language and aims to identify instances of therapy that are consistent or not with the Motivational Interviewing structure. For example, coded therapist behaviours include open or closed questions or affirmations ('great', 'thanks for coming') and patient behaviours include *change talk* ('when I stop drinking') or *sustain talk* ('I don't want to'). In each of the studies being described here, these codes were manually allocated to the text by up to three human raters.

Atkins et al., 2014 published the broadest piece of work on this subject that aimed to develop a text classification model that would be able to allocate MISC codes to future, unlabelled transcripts. Topic modeling methods were applied to a dataset of 148 motivational interviewing transcripts. This is a machine-learning approach in which topics, or clusters of words that co-occur in the text are automatically extracted from the corpus of documents. These can be extracted with or without the inclusion of specific labels that provide additional input that the model can learn from. In this case, the manually allocated MISC labels were included to assist the model in developing topics around these specific labels. The predictive performance of the developed model was then tested for each of the 12 MISC codes and assessed with Receiver Operating Characteristic (ROC) curves. These provide an indication of model performance in terms of sensitivity (true positive rate or the number of events identified over the total number present in the data) and specificity (true negative rate - the number of correctly identified non-events). The model was tested on its ability to provide labels both at the 'talk turn' (individual utterance) level and across the whole session. The results suggested that it performed poorly at the individual talk turn level but better at the session level. This suggests that the information contained in one utterance was not enough for the model to make a reasonable prediction. Additionally, the model performed better when predicting labels such as *questions*, *information giving*, or *reflections* that would be expected to have a more consistent semantic structure with c-statistics of approximately 0.8, than in the prediction of constructs that are more abstract such as *empathy*, in which the model performed less well with c-statistics around 0.7 (Atkins, Steyvers, Imel, & Smyth, 2014). It is not surprising that the codes associated with the most reliable linguistic structures would be best predicted by a linguistic model.

Three further pieces of work focused on specific codes within the MISC. A subset of 57 sessions of the data used in Atkins et al., (2014) was used to focus on one specific MISC code: *reflections* (Can et al., 2012). *Reflections* involve a therapist listening and returning to the patient what they have said either using the same or different words, often with the aim of guiding them through a particular problem. As opposed to developing topics as in the previously described piece of work, this research used 'n-gram' features as potential predictors within their model. In this case n-grams refers to a phrase of up to N consecutive words (where N is the utterance or talk turn length). This means that individual words and phrases were included as model features. Additionally, similarity features were developed. These were defined as instances in which an n-gram (word or phrase) was shared between a therapist utterance and the preceding patient utterance suggesting repetition of that phrase. As well as these language features, contextual information such as speaker identity was included in the model. The model subsequently developed aimed to identify whether or not an utterance contained a therapist reflection or not. The results suggested that the strongest model developed achieved an F-score of 80%. The F-score is a weighted measure of recall (also sensitivity or fraction of identified events over all events in the data) and precision (positive predictive value or fraction of identified instances that are relevant over all that were identified). 80% was considered a strong model by the authors but it is important to remember that it still leaves a large error margin. Though these results may be promising within a research context, they are too unreliable for clinical application.

The second piece of research working on a specific element within the MISC coding structure was published by Xiao et al. in 2015 and focused on labels of high and low therapist empathy. The study also included a project on automatic speech recognition but this will not be covered here. The empathy prediction research was carried out with 200 transcripts of motivational interviewing sessions that were manually labeled as containing either high or low therapist empathy. The features included in the model were slightly simpler than those described above and included individual words or short phrases of up to three words (or trigrams). The results of this model showed quite a strong performance of the automated detection of high or low empathy in therapy sessions with an overall F-score of 88.6%. For comparison, human coder agreement was estimated with an F-score of 90.3%, suggesting that the automated system was quite close to human performance. This is the highest performance to be found across the classification problems reported here.

Nonetheless, it is important to note that the model was validated using leave-one-out cross-validation, meaning that the model was estimated on all transcripts with the exception of those from one therapist and then the empathy level was estimated for those sessions using the developed model. This was repeated so that each session had an estimated binary empathy score. These results need to be replicated and externally validated, even more so because the therapy provided may be much more consistent within this data set than in general service provision given that the transcripts used came from two clinical trials of motivational interviewing.

The third piece of research looking at specific elements within the MISC was carried out by Tanana et al., in 2015 and focused on the identification of *change talk* and *sustain talk* in patient language. *Change talk* refers to patient language that indicates a willingness to change with regards to their harmful behaviour, in this case drug or alcohol abuse, and *sustain talk* refers to language that indicates resistance to change in the patient. Transcripts from 356 sessions of motivational interviewing that include the data used by Atkins et al. (2014) and Can et al., (2012) were used in this study. As was the case in the previous three studies described, manually allocated MISC codes were used as the gold standard or correct classification in the development of the model. The authors developed multiple model types and included the MISC code for the previous utterance as well as unigrams, bigrams and trigrams (one, two and three word phrases) as language features and predictors in the models. The details of the individual model types will not be included here. The results suggest that the strongest model developed in this data set to predict *change talk* achieved an F-score of only 22% and the best model to predict *sustain talk* achieved an F-score of only 24%. These results suggest very poor performance of these models in the prediction of patient motivation for change. It is clear that the approach followed in this study was not as successful as hoped. Some of this may be due to the difference between c*hange* and *sustain talk* being sometimes very subtle and therefore difficult to judge. For example, the phrase ‘I don’t need to drink’ may suggest *change talk* if the patient is convinced that alcohol consumption is a behaviour they can do without but may also be considered *sustain talk* if the patient is in denial about their need to take action. A different approach to solving the classification problem may be required in this case, such as including a different set of language features or different model structure.

The final piece of work within this section moves away from focusing solely on motivational interviewing and MISC codes towards considering multiple mental health treatment formats and the linguistic features that may allow a computer to discriminate between them. This piece of work by Imel et al., (2014) used the larger data set mentioned above of 1553 transcripts from a variety of mental health treatment approaches including medication management, cognitive behaviour therapy, psychoanalysis, motivational interviewing and brief relational therapy. The authors applied two types of topic modelling to the dataset. The first was what is known as unsupervised topic modelling in that no labels were provided to guide the model but it relied on extracting clusters of terms (topics) based on frequency and co-occurrence. This type of modelling was applied to extract 200 topics from the data set in order to explore it. These were then manually classified and labeled by the authors based on the terms they contained and organized into four areas: emotions, relationships, treatment and miscellaneous. For example, within the emotions area, the authors labelled five topics as anxiety, crying, hurt feelings, enjoyment, and depression. The anxiety topic contained words such as: ‘anxiety’, ‘nervous’, ‘panic’, and ‘tense’. These topics were developed to explore the data but also to develop a classification model that could discriminate between four types of therapy: Medication management, CBT, Psychodynamic and Humanistic/Existential. The overall prediction model performed with an accuracy of almost 87%, meaning that it classified a transcript correctly 87% of the time. In addition to developing this model, a second set of topic modelling was run but in this case the specific treatment labels (drug therapy, psychoanalysis, brief relation therapy etc.) were included so as to extract clusters of word that may be representative of individual treatment formats. However, these results appear to have only been exploratory in terms of observing the common terms and phrases associated with different therapy formats and the extracted topics were not included in the model described above.

A model that is able to discriminate accurately between treatment methods may be a useful step towards both automatic monitoring of therapy provision and identifying the specific active components of different treatment types. This work was put forward as an exploratory piece of research and, though the results leave room for improvement, they are promising. There is indeed potential for improvement with a range of further lingustic information that could be included, for example word order or multi-word phrases. The primary limitation of this study is in the data set used. Though it is sizeable and represents a diversity of mental health treatment formats, these are not evenly represented. For example, one case of psychoanalysis was represented by over 200 transcripts whereas a number of the medication management sessions may be only single sessions. Furthermore, sample transcripts provided by Carl Rogers and Albert Ellis were included that were up to half a century old. These are not necessarily representative of psychological therapy provision today.

In Imel et al. (2015), described earlier, a second problematic issue is evident with the labelling of the training data for the model. The labelling was not carried out following strict adherence to a manual or specific guidelines so reliability is uncertain, making the basis of the model weaker. A similar issue was present in the work by Tanana et al., (2015) whose classification results were the weakest of those presented. Their models were trained on manually annotated transcripts; however, the inter-rater reliability of these transcripts only just reached 61%, making it therefore difficult for an automated classification tool to reach high scores. The authors also note that they did not try all possible combinations in building the model, meaning that there is potential for different model structures to be developed and tested and that these may be more successful.

#### 2.4.1.7 Language used in outpatient consultations for individuals with a diagnosis of schizophrenia

The next section covers three pieces of research focusing on the language used in outpatient consultations for individuals with a diagnosis of schizophrenia or schizoaffective disorder (Howes, Purver, & McCabe, 2013; Howes, Purver, McCabe, Healey, & Lavelle, 2012a, 2012b). Though the datasets used in each publication do not appear to be identical, they all originate from a set of video and audio-recorded consultations between psychiatrists and patients attending assertive outreach and outpatient clinics. Assertive outreach teams work with individual with complex mental health needs who may need more intensive support than that provided by a community mental health team. The recordings were transcribed for analysis.

The first publication to be covered is a short paper based on a conference presentation. It focuses on *repair* and the association between *repair* patterns in outpatient consultations and patient adherence to treatment. *Repair* in dialogue can broadly be seen as a clarification or correction of what was said. It is the focus here as patterns of repair had previously been shown to be correlated with treatment adherence. A number of elements of *repair* were defined within the work. The focus was on ‘P2NTRI’ or ‘position 2 next turn repair initiator’, that is to say a phrase in which a speaker prompts another speaker to repair a previous utterance. The aim was to build a model to automatically detect this aspect of *repair* as well as a model to predict adherence to treatment. The linguistic features included in these models were a set of ‘high-level’ features defined within the study such as speaker identity, the number of words in each utterance, the number of backchannels (‘uh-huh’,’yeah’), the number of filler terms (a sound or word such as ‘er’ or ‘um’ that generally implies a pause in speech but indicates that the speaker has not finished their turn) or the number of portions of overlapping talk, for example. Additionally, patient unigrams (individual words) were also included.

The model predicting *repair* only achieved an F-score of 44% percent. But this low F-score was thought to be associated with the very low number of *repairs* of this type within the data (170 in 20,911 talk turns) and suggests that the model was able to identify almost half of these. The model predicting adherence reached an F-score of 70% with the model performing significantly worse (F-score of 35.5%) if only high-level features were included (no unigrams), suggesting that including the unigram features was crucial to model success. Though these results seem low, they appear to be preliminary or summary results and are associated with the work that will be covered next. Furthermore, this was a short paper that included only limited details about the methods applied, limiting understanding and the possibility to replicate the work. However, it has been assumed that a number of details provided in the next paper may apply here.

The next paper in question was a further piece of research on the language used by individuals with a diagnosis of schizophrenia in outpatient consultations with a psychiatrist. The aim here was to build and assess the performance of a number of classification tools that looked to predict symptom severity, the quality of patient experience and adherence. The linguistic features measured in this case were the same as those described previously, including a variety of high-level features as well as individual words. Patients were asked to complete the Positive and Negative Syndrome Scale (PANSS) (Kay, Fiszbein, & Opler, 1987), the Patient Experience Questionnaire (PEQ) (Steine, Finset, & Laerum, 2001) and clinicians rated patient adherence to treatment as good, average or poor. Each of these outcome scores was converted to a binary outcome with the boundary decided in order to achieve balanced groups on each side.

The results achieved by the classification tools in this case were very promising with models achieving close to 90% accuracy (overall percentage of correct classifications) for almost all outcomes, the exception being the communication subscale of the Patient Experience Questionnaire for which 80% accuracy was reached. These results are much stronger than those presented in the previous publication. However, as with the previous publication, a large proportion of the success in these models appears to be attributed to the inclusion of individual words, or unigrams, as predictors in the model. Without these, model accuracy falls to around 50% in most cases. Furthermore, it was noted by the authors that there was little overlap in the sets of individual words that were predictive of different outcomes. For example, there was only one common word (‘mates’) in the word features selected as predictors of overall Patient Experience Questionnaire outcome and adherence. This can be seen to support the idea that individual outcomes may be best predicted by specific, tailored models.

There are two primary limitations to the work described above. The first is that given the small number of transcripts (131) included in the study, the reliance of the model on individual words may lead to a risk of the models developed over fitting this particular dataset. That is to say that the model would be describing random error, paying attention to random noise rather than measuring the signal. This makes it less generalisable or applicable to other data sets. It would therefore benefit from further testing on a larger, independent data set. Secondly, the outcomes were reduced to binary measures in the classification tasks. Though this may ease model development and improve performance, there is a loss of information as compared to a continuous outcome. Nonetheless, the results in this study are promising in a young field.

Finally, one important limitation of a number of machine learning models of language is the difficulty in interpreting these as they are often ‘black box‘ mechanisms or include predictors that are difficult to attribute meaning to. For example, the inclusion of individual words in the model developed improves the performance of this model but the relevance of individual words is very difficult to interpret. This is a problem that Howes et al., (2013) attempted to solve in their next piece of work in this area.

This next piece of research was based on the same type of data, though not an identical set. It was made up of 138 transcribed records of outpatient psychiatric consultations from patients with a diagnosis of schizophrenia or schizoaffective disorder. The same outcome measures as were recorded in the previous study were recorded here with the addition of the Helping Alliance Scale (Priebe & Gruyters, 1992) as a measure of the therapeutic relationship. Twenty hand-coded topics were developed to fit the data set and labels for these applied to each segment of a consultation. Examples of these topics were *medication*, *physical health*, and *coping strategies*. In terms of computerised linguistic analysis, the primary difference between this piece of work and the previous two studies was the method of linguistic analysis applied. In this case, topic modelling methods were used, an approach that has been applied in previous work in this review. The model was set to extract 20 topics from the textual data and no labels were included to guide the model. These twenty topics were then manually evaluated by two different groups of human raters in order to interpret them and assign descriptions to each topic. The automatically extracted topics and hand-coded topics were then compared and correlated to assess any associations between them. The results suggested that there were some strong associations between hand-coded and automatically coded topics such as between *medication* (hand-coded) and *medication regiment* (automatic) or between *alcohol, drugs and smoking* (hand-coded) and *substance use* (automatic). All significant correlations reported were positive, suggesting that there was some similarity between the topics extracted by hand through qualitative analysis and those extracted automatically.

The next stage of analysis in this paper involved the development of predictive models. In a first instance, correlations between both sets of topics and scores on the PANSS were considered and a number of significant positive associations were measured. It was found that general and positive symptoms on the PANSS were significantly correlated with both the hand-coded and automatically coded versions of the *psychotic symptoms* topics. The agreement between hand-coded and automatically extracted topics, where it was present, supports the meaning attributed to the extracted topics. This study also included a number of classification experiments similar to those described in Howes et al., (2012b), with binary outcomes for the measures of symptoms, patient experience, therapeutic alliance and adherence. Among these classification experiments, the best results were obtained for the prediction of clinician rated therapeutic alliance with a model including hand-coded topic information and patient and clinician gender and identity information. This model achieved an overall accuracy rate of 75.8%, meaning that the model was correct 75.8% of the time. This model did, however, rely on the inclusion of clinician identity factors. Based on the automatically extracted topics alone, this model achieved 65% accuracy, a gain of only 15% accuracy over a random model that would be expected to achieve 50% accuracy. The success of models predicting binary scores of other outcomes was generally lower, achieving over 60% accuracy in only a handful of cases: adherence, the measure of communication barriers within patient experience, PANSS general symptoms and PANSS general symptoms. However, it seems apparent that the hand-coded and automatic topics perform differently, with automatic topics performing better in predicting adherence and hand-coded topics performing better in the prediction of symptoms.

Though the performance of the classification models in this last piece of work leaves large room for improvement, the use of topics allows for easier and more comprehensible interpretations of any associations found. This stands in contrast to the results of Howes et al. (2012b) above in which the models performed better but where the features included were individual words, making interpretation of the results difficult. This last study was put forward as an exploratory piece of work on the application of topic modelling to this data set and the authors concede that the method applied was relatively simple and that more complex forms of topic modelling are available that may lead to improved performance of the classification models. Together, the three pieces of work described in this section demonstrate a number of approaches to linguistic analysis and in particular to the different types of linguistic features that can be extracted and included in classification models. This ranges from high level features such as patient and clinician identity or numbers of questions, affirmations or utterances in a session to the inclusion of individual words as features or the development of topics based on the distribution and co-occurrence of words throughout documents. In this data set it appears that the least interpretable models were the most successful in terms of classification accuracy but the exploration of topics in the sessions leads the way towards greater insight as they are either hand-coded or attributed meaning by human coders. In machine learning work where a large number of factors are considered, data sets are usually much larger than those explored here so it would be important to use the information gathered through these studies and explore and test the emerging models on a larger data set.

#### 2.4.1.8 Topic modelling work on cognitive behaviour therapy data – Howes et al., 2014

The final piece of work that will be described in this review was carried out on a dataset that is highly relevant to the present investigation. The data studied consists of approximately half of the online cognitive behaviour therapy data that is the focus of the research work contained in this thesis. These are 882 transcripts from 167 patients with a variety of mood or anxiety-based mental health issues attending therapy (provided by Ieso Digital Health) carried out online over an instant messaging platform (Howes et al., 2014). Two mental health outcomes were recorded for each session: the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder Scale (GAD-7). As these two measures are highly correlated, only the PHQ-9 score was considered and this was turned into two binary measures. Firstly, as a high/low score of PHQ-9 and secondly, as measure of change between a pre-treatment PHQ-9 score and the current score, the binary outcome on this last measure was of improvement versus no improvement.

In terms of the linguistic analysis and measures used from this dataset, three features or sets of features were extracted. The first was sentiment, measured using a computerized sentiment analysis approach called Sentimental (Purver & Battersby, 2012), developed by one of the authors. Scores between -1 and 1 were attributed to transcripts, with negative scores associated with negative sentiment and positive scores associated with positive sentiment.  A measure of anger was also automatically extracted with the same program. The main approach, however, was unsupervised topic modelling. This is the same method that was applied in Howes et al., (2013) above and allows the automatic extraction of topics in text based on the distribution of words within it. Correlation analyses were carried out to determine the associations between these language measures and the outcome scores and found a significant association between mean sentiment score and PHQ-9 score as well as between the number or words used by a therapist and PHQ-9 score. Additionally, two topics were found to be significantly associated with PHQ-9 score. However, no significant associations were found with the PHQ-9 change score. The first topic found to be significantly correlated with outcome was positively associated with PHQ-9 score and included words such as ‘gp’, ‘depression’, ‘help’ or ‘therapy’ that may suggest a common theme but also words such as ‘make’, ‘today’ or ‘little’ that are less clearly connected to the first set of words. Though there may seem to be a common subject in some topics that the terms in these topics can be attributed to, this is not always the case.

Classification tasks were also included in the analysis. The aim was to predict PHQ-9 outcome (score of above or below 10) and PHQ-9 change scores (improvement or not), both as binary outcomes. For these tasks, the authors included individual words and n-grams (multiword phrases) as language features. In terms of predicting outcome, the best model using language features alone (as opposed to patient and therapist identity information) performed with an F-score of 71% with a combination of sentiment measures and topics as included features. This combination of features did not perform as well when predicting the change score, however, and the results suggest that a model using n-grams (word and short phrases) performed better on this task, reaching an F-score of almost 70%. These results were put forward as promising in terms of model performance but when considering the potential clinical applications of these, this level of accuracy would not be high enough for implementation in clinical practice.

Determining what would be accurate enough to suggest clinical implementation is difficult, as it depends on whether the tool is used for monitoring or diagnostic purposes and whether it is to be used as a sole indicator or part of a battery of measures. However, research suggests that the PHQ-9 had a specificity and sensitivity of 88% for major depression when validated on a data set of 6000 patients across 8 primary care clinics and 7 obstetrics-gynaecology clinics (Kroenke, Spitzer, & Williams, 2001). I would therefore suggest that a computerized or statistical tool looking to identify the presence of depression through the analysis of language should reach similar and ideally improved rates of sensitivity and specificity prior to being considered for clinical application. Accuracy rates reaching 90% would be preferable. Similarly, the GAD-7 was found to reach a sensitivity rate of 92% with 8 points used as a threshold for diagnosis but only 76% specificity. Increasing the threshold will increase specificity and lower sensitivity (Spitzer, Kroenke, Williams, & Löwe, 2006). Given these scores associated with the GAD-7, a similar accuracy rate of 90% or more in a computerised tool to detect anxiety would be worth considering for clinical application.

Based on the research on outpatient consultations with individuals with schizophrenia described above, it may be useful to consider the development of hand-coded topics as well as the automatically extracted ones, or to modify the extracted topics so that the words contained within them are more consistent with that theme and descriptions allocated. It is possible that, as was the case with the work on schizophrenia, hand-coded topics may help in the prediction of symptom-based outcomes such as the PHQ-9.

#### 2.4.1.9 Brief discussion of classification model approaches

The research in this section mostly includes work around two broad themes, with one paper not fitting into either of these. The first subset focused on identifying and recognizing specific behaviours in textual data that were characteristics of the therapy format being applied. These were primarily behaviours defined within the motivational interviewing skills code such as the use of *reflections,* therapist empathy and patient language that provides evidence of willingness to change or not (Atkins et al., 2014; Can et al., 2012; Howes et al., 2012a; Tanana et al., 2015; Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015). The second subset focused more on mental health outcomes or measures of therapeutic experience such as symptom scores or scores of therapeutic alliance or adherence (Howes et al., 2013; Howes, Purver, & McCabe, 2014; Howes et al., 2012b). With the exception of the paper by Tanana et al., (2015) that sought to identify change and sustain talk in patient language, the classification models in the first subset for identifying specific behaviours in text performed better than those developed to predict mental health outcomes or measures of patient experience and adherence.

Within the research in this subsection, there were two data sources that appeared to be studied from a variety of angles and with the application of different methods in terms both of the linguistic features extracted and the prediction models developed. This seems to be an approach that is common within computational linguistics in which the same, or a very similar data set is studied by different research groups, or the same research group multiple times so as to apply a variety of methods, presumably in order to determine the best approach for the given data format. This approach serves as a form of discovery that appears to be symptomatic of a field in which it isn't clear which approach will work best for different data formats and outcome measures. Additionally, as was the case in the first section of this review, a major limitation in a number of the research studies presented here is in the data set used. The application of machine learning methods and the development of classification models with large numbers of predictors generally require a large dataset in order to perform sound analysis and avoid developing a model that will not generalize to other data. Though some datasets described do contain large numbers of transcripts (Imel et al., 2015), others are smaller and may run this risk (Atkins et al., 2014; Can et al., 2012). Understandably, much of the work presented here is exploratory and requires replication and validation on larger, more diverse datasets in the future.

**Table 2-2 Summary tables of methods and outcomes from studies selected for review.**

| Citation | Language variables (if applicable) | Qualitative/Manual method (if applicable) | Outcome measures | Analysis (statistical) | Main findings |
|---|---|---|---|---|---|
| Anderson, et al., (1999) | High and low affect, CALAS verb categories (stative, action or process), and stylistic complexity | Minor corrections during transcription | Adherence and presence/absence of repair | 2x2x2 mixed design MANOVA | Significant overall three way interaction. Differences in therapist verb use in high affect segments between good and poor outcomes |
| Atkins et al., (2014) | Extracted topics based individual words and word phrases | Motivational Interviewing Skills Code (MISC) labels | MISC code labels | ROC curves and Cohen's Kappa | Higher reliability in codes with more reliable semantic structure. Strongest models for Open Questions (0.81) and reflections (0.8) |
| Can et al., (2012) | N-grams and similarity features | Manual coding of reflections in therapist language using MISC codes | Reflections as measured by MISC | Classification experiments | Models using context meta-features and n-grams performing best (F-score of 0.80). |
| Fontao et al., (2008) | Measures of emotion and abstraction language | 16-item scale for group experiences. Therapeutic cycles | Therapeutic factors from Kiel Psychotherapy Process scale. Two main factors considered: group interaction and therapeutic process | MANOVA and independent samples t-tests | Therapeutic factors and group interaction scores differ across four language patterns/phases of a cycle |
| Haug et al.,(2008) | Indegree, Outdegree, activity and 52 LIWC variables | | 1. Group Evaluation Questionnaire (GEQ)<br>2. Group Relationship Questionnaire (GRQ) | Correlation analyses | Other group members (ID) play greater role in satisfaction with therapy than therapist (IDT). Some significant correlation between language features and GEQ and GRQ. |
| Howes et al., (2012a) | High-level language features (e.g. number of words, questions etc.) | Manually annotated instances of repair | 1. Positive and negative syndrome scale (PANSS)<br>2. Patient Experience Questionnaire (PEQ)<br>3. Adherence classified by clinician | Classification experiments | Models including lexical features showed much better performance than using high-level features alone. Best models reached over 90% accuracy. |

| | | | | | |
|---|---|---|---|---|---|
| Howes et al., (2012b) | High-level language features (e.g. number of words, questions etc.) | Manually annotated instances of repair | Adherence and presence/absence of repair and comparison with model. | Classification experiments | Classification model of repair reached 44% F-score and adherence classification reached an F-score of 70% |
| Howes et al., (2013) | Automatically extracted topics | 20 hand-coded topics | 1. Positive and negative syndrome scale (PANSS) 2. Patient Experience Questionnaire (PEQ) 3. Adherence 4. Helping alliancs Scale (HAS-D) | Correlations and classification experiments | Some topics were correlated with symptom scales. Best classification models were obtained for HAS score (75%). |
| Howeset al., (2014) | Topics, sentiment measures, high-level features, n-grams | | Binary measures of PHQ-9 outcome and change score | Correlations and classification experiments | Best models in classification experiment reaching around 70% accuracy. |
| Imel et al., (2014) | Extracted topics | | Therapy type | Multiple classification experiments | Machine learning models to discriminate between therapy types performed with cross-validated error rate of 13.3 %. |
| Tanana et al., (2015) | N-grams and word vectors | MISC codes of change and sustain talk | MISC codes of change and sustain talk | Classification experiments | Best models reached 0.22 and 0.24 for classification of Change and Sustain Talk respectively. |
| McCarthy et al., (2014) | Emotion-Abstraction patterns | | High or low improvement. | Analyses of variance | Significant differences in time spent in connecting and relaxing patterns between most and least improved patient groups. |
| Xiao et al., (2015) | | Therapist empathy ratings (high or low) from MISC. | High or low empathy categorisation - MISC code labels. | Classification experiments | The automated labelling system reached 85% accuracy for binary prediction of empathy. |
| Van der Zanden et al., (2014) | LIWC variables. | | 1. Depression symptoms (CES-D) 2. Anxiety symptoms (HADS-A), 3. Perceived control (Mastery Scale) | Correlation and regression analyses | Significant associations between client word use and treatment outcome and adherence. |

## 2.5 Discussion

This review has shown that a number of approaches to linguistic analysis of therapeutic dialogue in mental health research have provided insight into therapy provision and course. The most striking point to make concerns the diversity of methods and approaches that have been applied. In the last few years a greater proportion of work has emerged from or in collaboration with the field of computational linguistics where there appears to be a preference for the development of models to perform classification tasks. In these, the focus is on accurate prediction rather than interpretation of specific language features and their association with a given outcome (Can et al., 2012; Tanana et al., 2015).

The collaboration of mental health and computational linguistics fields can lead to the development of topics that are interpreted within the mental health context as well as applied in classification work (Howes et al., 2013). The research work emerging from the mental health field tends to focus more on how linguistic features relate to mental state and the interpretation of these in terms of what they mean about the patient or therapist. In these cases, prediction accuracy seems to be less of a concern and the statistical analyses focus on measures of association (Haug, Strauss, Gallas, & Kordy, 2008; Van der Zanden et al., 2014).

The case for collaboration across disciplinary fields seems clear in order to make the most of current technological capacity and understanding of mental health. There is also a great deal of overlap in the data used by a number of the papers included in the review, as well as the use of public, perhaps outdated, sample psychotherapy sessions (Atkins et al., 2014; Imel et al., 2015; Tanana et al., 2015; Xiao et al., 2015). This further highlights the need for collaboration on a more practical level in terms of access to relevant and diverse datasets.

The second conclusion of this review is that when considering individual measures of association between language features (where these are discussed) and outcome scores or measures relating to psychological therapy, results such as correlation analyses are often relatively weak despite being statistically significant (Haug et al., 2008; Van der Zanden et al., 2014). This suggests that these language features are providing an element of insight into mental states but cannot necessarily be seen as directly representative of them and that the relationship is perhaps more complex than originally thought. Two elements that are not mutually exclusive could be at play here. Firstly, modelling or understanding mental state based on language features may require the combination of a range of language features, including language features for which the association with mental health outcomes has not yet been researched. Secondly, individual variability potentially plays a large part in how an individual expresses his or her psychological state, suggesting that this association may well be mediated by or interact with non-linguistic personality factors. This would mean that these affect language choice and use and that greater understanding of how these influence verbal or written expression may lead the way towards more precise predictions of mental health outcomes.

Conversely to what is described in the last paragraph, results from classification work looking to determine quite narrow features were generally more successful. This may be due to the language used being quite closely associated with the features by definition. This is the case for *reflections* or *questioning*, for example, where the sentence structure is quite rigid and more predictable as well as being directly associated with a particular therapeutic tool or skill (Can et al., 2012; Xiao et al., 2015). The gap a model needs to breach between the language used in an utterance and determining whether it is a question or not is smaller than that between the language used by an individual and determining their symptoms or mental health outcome. This may explain why the work focusing on classifying the presence of very specific features seems to have stronger results than that looking to model therapy outcome scores. It may be, however, that the narrow features described can be seen as features that will then be suitable as predictors within a broader model predicting outcome. The prediction of mental health outcomes is a prospective challenge, whereas identifying expressions of specific behaviours in text is a cross-sectional, immediate task.

A number of the papers covered in this review put forward specific language features with associated definitions that allow for automated measurement of these features. For example, measures of *indegree* and *outdegree* (Haug et al., 2008) or relative emotion-abstraction patterns (Fontao & Mergenthaler, 2008; McCarthy et al., 2014). This development of original language features and the sharing of these within the research community would allow the same features to be measured automatically, and thus objectively, across different data sets and research projects. This kind of replication should be encouraged as it will boost the external validity of results.

### 2.5.1 Limitations of review

The limitations of this review primarily concern the variety of studies included and the problems this poses in terms of drawing conclusions. Though the review was carried out to inform a project working on language in online cognitive behaviour therapy, there were too few studies to restrict this review to language studied within only this context. Therefore, it was expanded to include written textual data from therapeutic dialogue. The mixed disciplines involved in this type of research and variety of methods applied make it difficult to perform an informative meta-analysis.

The databases searched were primarily health and life sciences related, meaning that a number of the papers included here were not found in the main database search. In 2013 a workshop on ‘Clinical Psychology and Computational Linguistics;, initiated by the North American chapter of the Association for Computational Linguistics, focused specifically on the application of computational linguistics methods in a mental health context. This prompted the inclusion of the Association for Computational Linguistics’ anthology in the search and the hand-search of references in relevant papers increased the number of relevant studies found but it is possible that there is further work not published within the expected channels that was not found through this search. With the exception of the archives of the Association of Computational Linguistics, the search focused on mental health research sources, as this is the background of the project. Expanding to wider databases and including others with a greater focus on computational research may be helpful and reveal further relevant research.

### 2.5.2 Implications

The spread of research approaches and low levels of replication may be a symptom of a young but rapidly developing field, but may also be associated with the way this field of research sits across multiple disciplines. The majority of the more technically advanced work has been carried out within the field of computational linguistics, and the developed models sometimes have limited application to clinical reality in the state in which they are published. Furthermore, in some cases the work appears to be limited in its access to current clinical data as a number of studies employ the same datasets repeatedly as well as using old, possibly outdated recordings of therapy sessions (Atkins et al., 2014; Can et al., 2012; Imel et al., 2015). On the other hand, more clearly applicable work is emerging from research being carried out in a more clinical context but the methods applied may not be the most computationally advanced (Haug et al., 2008; Van der Zanden et al., 2014). It is important to bear in mind that the latest technology may not always be the best; one major recommendation for this field of research is to promote collaboration between computational linguists and mental health academics and professionals. A multidisciplinary approach with the best expertise might be most likely to bridge gaps between research fields. Where this has been done, interesting results are found (Howes et al., 2013, 2014).

As the field is growing rapidly, it is important that a new review of research be carried out in the future. Given the increasing rates of publication in this field, a future literature review that incorporates a wider field of study (as suggested in the previous section) is likely to find increased numbers of relevant papers. If the numbers allow it, it may be advisable to narrow the review question further in order to focus either on a specific form of psychological therapy or a specific analysis method. This would therefore allow more direct comparisons to be made between research studies and provide robust conclusions about the value of a given method of linguistic analysis or the value of linguistic analysis within a given type of therapy.

This literature review confirms that very little work has been carried out on the specific type of data that is the focus of the current research project, namely, online text-based and one-to-one cognitive behaviour therapy. The therapy format itself is a recent development and is likely to be rare. There are a number of online befriending, counselling and mental health services but these do not appear to have been, as yet, the focus of research into language use. This review also suggests that no work has been done to consider the potential of using text mining methods in working with this or a similar type of data format. Text mining differs from most of the methods described here as it involves the development of language features through an interactive and iterative process working with the data at hand. Some of the approaches included in this review would therefore be applicable within text mining, such as specific language dictionaries (e.g, abstraction and emotion or LIWC) or the grammatical relationships put forward by Anderson et al. (1999). It is feasible that features developed through text mining could be incorporated into machine learning algorithms to determine ideal weighting of these in a model. Therefore, text mining does not necessarily sit in opposition to the methods presented here, but may provide a helpful tool or step in an analytical process that has not yet been fully exploited.

The clinical applications of the research reviewed here are currently limited, though they may have considerable potential. The development of classification tools able to determine the presence of selected features in therapeutic dialogue such as repair, the clarification or correction of a phrase, in outpatient consultations (Howes et al., 2012a) or the presence of empathy in therapist language (Xiao et al., 2015) may be of particular interest in monitoring therapy practice and learning about active ingredients in mental health treatment. There is potential in feature detection of this type despite the need for further work in order to develop more consistent, generalisable and context-adaptable tools. Diagnostic classification of individuals based on language use seemed less successful where it was attempted in the work reviewed here (Howes et al., 2013) but this can be expected due to the more complex nature of the classification and factors contributing to a diagnosis. Breaking down elements of diagnosis into more manageable classification tasks may be a way forward in this area. This would open up a whole range of possibilities for including language use in the diagnostic process with the potential of an objective, automated ‘second opinion’ that might assist clinicians in their work.

## 2.6 Conclusion

Computerised analysis of language in psychological therapy is an area that has generated considerable interest in the past three years. Most of the work reported on here is very recent. The work is also still experimental and exploratory in nature with a range of methods and linguistic features being considered as potential candidates for analysis but only limited evidence of replication or building on previous work. No work has been done looking at the value of text mining in this form of research and very little research has looked at data from online text-based individual therapy, leaving a gap to be filled for the research that will follow in this thesis.

The research work that has been carried out is nonetheless promising in showing how best to use language to understand and improve psychological therapy. There is undeniable value in being able to carry out detailed analysis of language in a therapeutic setting to gain meaningful insight through observational means when the volume of relevant data is ever-increasing. It can provide additional insight into the therapeutic experiences to what can be recorded in a questionnaire. A questionnaire will normally focus on a pre-defined area and can guide the focus of an individual’s responses, whereas a natural language record of what is said in a treatment session provides direct evidence of their therapeutic process, such as how they express their difficulties, respond to suggestions from a therapist and react to successes and setback over the course of treatment. This language may include evidence of immediate reactions as opposed to a potentially delayed response to a questionnaire or interview. This could provide greater nuance, in both language use and content, than answers to specific questionnaire items. However, patient and therapist comment about their experience of therapy can also provide important context to the language used within treatment sessions. The analysis of language in treatment can therefore be seen as complementary to other research methods that consider patient behaviour, measures of outcome, and patient experience of treatment.

It seems that there is as yet no clear direction for clinical application of this kind of work as the potential applications and methods by which to do this vary so widely. However, as more work is generated in the field, the value of specific methods or linguistic features for a particular application, such as predicting outcome or determining the presence of particular therapist qualities, may become apparent. There is no doubt that the activity in this field is likely to grow as technical skills develop and more mental health work is computerized.

# Chapter 3. Methods

This chapter will cover the methods for the entire research project reported on in this thesis. Following the description of the participants and data sample, a full description of the linguistic methods used can be found. The work was carried out in stages, with each set of linguistic features, extracted from the text and associations with outcomes explored in statistical models in turn, thus the work is not presented here in strict chronological order. A description of the statistical analyses applied will follow the linguistic methods as, though the language features were tested separately, the same process was followed for statistical analysis for each set of features, with a final model combining relevant variables from each set of previously tested features.

Ethical approval for this project was obtained through the proportionate review sub-committee of the NRES Committee London – Riverside. The Research Ethics Committee (REC) reference is 13/LO/1929 and IRAS project ID is 141708.

## 3.1 Data

The data used in this study were two sets of transcripts, a development and a validation set, from online text-based cognitive behaviour therapy delivered by Ieso Digital Health to patients who have been referred within the NHS by their General Practitioner (GP). Development set is here used to refer to the data set with which associations between linguistic features were explored and predictive models were first fitted and developed. The validation set refers to a second data set, independent of the first, on which statistical models were tested and therefore externally validated. A further dataset, transcripts from the IPCRESS trial (D. Kessler et al., 2009) (see section 3.1.1.4), was also used to assist development of some language features.

### 3.1.1 Participant groups

#### 3.1.1.1 Ieso Digital Health online therapy

Ieso Digital Health are the largest provider of online CBT in the United Kingdom. They provide online CBT on behalf of the NHS within the context of the Improving Access to Psychological Therapies (IAPT) initiative. One-to-one, text-based online cognitive behaviour therapy is provided over a purpose-built instant messaging platform. This means that the therapy is carried out with both therapist and patient present online simultaneously and follows the same structure as face-to-face treatment, but all communication is typed. Patients are referred by their GP and are allocated a therapist based on their provisional diagnosis, therapist availability and expertise. IAPT works within a stepped care framework, where the ‘step’ is associated with the severity of an individual’s mental health condition and refers to the type of care they will have access to. GP contact puts patients at Step 1 and patients are assessed by their therapist and allocated to the appropriate step for their mental health needs. Ieso works with patients allocated to Step 2, Step 3 and Step 3+, where Step 3+ refers to anyone above a step 3. The allocated step will have an impact on how many sessions of psychotherapy a patient is offered. According to the IAPT service specification, patients are offered between six and eight sessions on Step 2 and eight sessions or more, sometimes up to twenty, on Step 3 or above (Department of Health, 2011).

Patients are considered by the service (Ieso Digital Health) to have completed treatment when they are discharged upon agreement with their therapist. If they have not been discharged and do not return for treatment, they are considered to have dropped out. In some cases it is established during assessment or early treatment sessions that the service is not the best option for a patient and they are referred elsewhere or back to their GP. Though the service is primarily designed for individuals with anxiety or depression diagnoses, they do work with individuals with a range of other provisional diagnoses.

#### 3.1.1.2 Development data set

The development set is made up of transcripts and associated outcome and demographic information for 661 individuals who were referred for online therapy between May 2013 and April 2014. 233 of these completed treatment, 132 were in treatment at the time of data collection and 218 dropped out for a combination of reasons. These could be due to personal preference, their unsuitability for this particular service and referral elsewhere and other undisclosed reasons. 78 patients never activated their account to begin treatment. Combining all patients who attended a session this made for a total of 2552 transcripts. Assessment sessions and short sessions are 30 minutes long and full sessions are 60 minutes long. This means that there is great variability in the length of transcripts and number of words typed. The development set contains 451 women and 208 men (information was not available for 2 patients). Tables for age groups, provisional diagnosis, step group and patient status are included below.

**Table 3-1 Patients by age group in development set.**

| Age group | Frequency | Percent |
|---|---|---|
| **Under 18** | 1 | 0.2 |
| **18 – 29** | 208 | 31.5 |
| **30 – 40** | 203 | 30.7 |
| **41 – 50** | 147 | 22.2 |
| **51 – 60** | 77 | 11.6 |
| **Over 60** | 24 | 3.6 |
| **Not known** | 1 | 0.2 |
| **Total** | 661 | 100.0 |

**Table 3-2 Patients by diagnosis in development set**

| Diagnostic group | Frequency | Percent |
|---|---|---|
| **Anxiety** | 184 | 27.8 |
| **Depression** | 283 | 42.8 |
| **Eating Disorders** | 2 | 0.3 |
| **Stress** | 6 | 0.9 |
| **Obsessive Compulsive Disorders** | 16 | 2.4 |
| **Somatisation** | 14 | 2.1 |
| **Sexual Disorders** | 1 | 0.2 |
| **Mixed anxiety and depression** | 66 | 10 |
| **Other diagnoses** | 45 | 6.8 |
| **No provisional diagnosis given** | 44 | 6.7 |
| **Total** | 661 | 100 |

The category of ‘other diagnoses’ includes the following: adjustment disorders, irritability and anger, mental disorders, not otherwise specified, and problems in relationships. These are groupings provided by the service based on GP and triage assessment.

**Table 3-3 Patients by Step in development set**

| Step | Assessment attended | Frequency | Per cent |
|---|---|---|---|
| **Step 2** | - | 113 | 17.1 |
| **Step 3** | - | 277 | 41.9 |
| **Step 3+** | - | 91 | 13.8 |
| **None allocated** | Yes | 18 | 2.7 |
| **None allocated** | No sessions attended | 162 | 24.5 |
| **Total** | | 661 | 100.0 |

#### 3.1.1.3 Validation set

The validation set is made up of transcripts and appointment information for 376 individuals who were referred for treatment between July 2014 and April

2015. 185 individuals completed treatment, 171 dropped out of treatment after starting the course, 18 did not complete treatment as they were found to be unsuitable for the service and 2 were referred back to their general practitioner. There are 1667 transcripts in the validation set. The group whose data make up the validation set was made up of 279 females, 96 males (gender not disclosed for one individual). Tables for age groups, provisional diagnosis and step group are included below.

**Table 3-4 Patients by age group in validation set**

| Age group | Frequency | Per cent |
|---|---|---|
| **18 – 29** | 117 | 31.12 |
| **30 – 40** | 100 | 26.6 |
| **41 – 50** | 95 | 25.27 |
| **51 – 60** | 43 | 11.44 |
| **Over 60** | 21 | 5.60 |
| **Total** | 376 | 100.0 |

**Table 3-5 Patients by provisional diagnosis in validation set**

| Diagnostic group | Frequency | Per cent |
|---|---|---|
| **Anxiety** | 75 | 19.9 |
| **Depression** | 52 | 13.8 |
| **Eating Disorders** | 3 | 0.8 |
| **Stress** | 5 | 1.3 |
| **Obsessive Compulsive Disorders** | 8 | 2.1 |
| **Somatisation** | 5 | 1.3 |
| **Sexual Disorders** | 2 | 0.5 |
| **Mixed anxiety and depression** | 97 | 25.8 |
| **Other diagnoses** | 43 | 11.4 |
| **No provisional diagnosis given** | 86 | 22.9 |
| **Total** | 376 | 100 |

**Table 3-6 Patients by step in validation set**

| Step | Frequency | Percent |
|---|---|---|
| **Assessment attended but no step allocated** | 26 | 6.9 |
| **Step 2** | 76 | 20.2 |
| **Step 3** | 251 | 66.7 |
| **Step 3+** | 23 | 6.1 |
| **Total** | 376 | 100.0 |

#### 3.1.1.4 Differences between data sets

The demographic variables set out in the tables above put forward some differences in the diagnostic profiles of the two populations providing the data for analysis within this project. In terms of provisional diagnoses, the validation set saw a larger spread of diagnoses with a high number of patients presenting with mixed anxiety and depression or other mixed diagnoses. In contrast, the development data set had a majority of depression diagnoses (over 40%) as compared to the 14% of depression diagnoses in the validation data set. There was also a small difference in the spread of allocated step, providing an indication of severity of mental health disorder. A greater proportion of patients were allocated to Steps 3 and 3+ in the validation set suggesting a population with higher severity of mental illness. The large portion of patients in the ‘assessment’ category in the development set is associated with the higher drop-out rate in this data set as it indicates that patients dropped out prior to completing an assessment session with Ieso Digital Health. Finally, there are only slight differences in the age profile of the two populations with a slightly larger spread in age group in the validation set including more patients over 60 than were found in the development data set population.

The differences in geographical location of the patient populations are not shown in these tables. The two data sets contained data from patients in two different areas of the South of England. It is possible that there are socio-economic and educational differences between these populations but these were not measured in this dataset.

#### 3.1.1.5 IPCRESS data

The IPCRESS trial (D. Kessler et al., 2009) was carried out between 2006 and 2009 and aimed to determine the effectiveness of online text-based cognitive behaviour therapy. During the trial, 297 individuals were either allocated to online therapy or face-to-face CBT, which involved spending time on the waiting list, but all were eventually offered a course of cognitive behaviour therapy. A random sample of approximately 20 transcripts were read and used as a guide for the development of some of the linguistic features and to provide the author with further understanding of the therapeutic process. Demographic and outcome information for these patients was not available and only anonymised transcripts were accessed.

### 3.1.2 Data format

The development and validation set were provided as two large files, one for each data set, containing date and reference number for each session, and time and speaker information for each message sent. Two different methods of anonymisation were applied to the development and validation sets, both prior to transfer of the transcripts from Ieso Digital Health.

Two spreadsheets accompanied each dataset. These included case and appointment information. Case information contained the demographic details for each patient (anonymised) along with their provisional diagnosis, completion of treatment status and step allocation. The appointment information contained details of attendance, length of appointment, time and date information for each appointment as well as outcome scores that patients were requested to complete prior to each session.

### 3.1.3 Outcome scores

Throughout treatment, patients are requested to complete various questionnaires and scales depending on their provisional diagnosis. However, all are required to complete the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder Scale (GAD-7) in accordance with the IAPT Outcomes Framework (‘IAPT Data Handbook’, 2011). Patients are requested to do this up to two days before a therapy session. These are the outcome scores that were used throughout this project.

#### 3.1.3.1 Patient Health Questionnaire (PHQ-9)

The PHQ-9 scale is a nine item self-report questionnaire used to assess levels of depression. The individual completing the scale is asked to rate on a four-point scale, ranging from ‘not at all’ to ‘nearly every day’, how often they experience a particular symptom associated with depression. Each item refers to a different DSM-IV related criterion for depression, which could be low mood, change in appetite or loss of motivation, for example. The scores were considered as continuous measures of depression outcome.

#### 3.1.3.2 Generalized Anxiety Disorder Scale (GAD-7)

The GAD-7 is a seven item self-report questionnaire used to assess levels of anxiety. Similarly to the PHQ-9, an individual is asked to rate how often they have been affected by a set of common signs of Generalized Anxiety Disorder on a scale ranging from ‘not at all’ to ‘nearly every day’. The items include symptoms of anxiety such as feeling restless, feeling nervous or anxious or finding it difficult to relax.

## 3.2 Materials

This project relies primarily on I2E, a text mining platform through which linguistic features are both developed and extracted in order to provide numerical data for analysis. I2E was developed and is provided here by the second commercial partner associated with this project, Linguamatics Ltd.

Further detail on the use and application of I2E within this project can be found in the following section 3.3. Data editing and checking was carried out using Excel by Microsoft Office. For the statistical analyses, the statistical package STATA version 12.0 has been used throughout the research project.

## 3.3 Linguistic analysis methods

### 3.3.1 Text mining with I2E

I2E by Linguamatics was used to extract linguistic features from the textual data. I2E is a specialised software that provides facilities to search large quantities of textual information using manually built search phrases, called queries. Natural Language Processing (NLP) methods are applied by the software to best exploit the data and capture the relevant information or detail in the text. These include processes such as stemming; reducing words to their stem by removing suffixes or inflections and parts of speech tagging; determining the grammatical role of a word (noun, verb, object, etc.) within a phrase. These methods essentially allow the software to ‘read’ the input textual data.

For the purposes of this project, the most important aspect of the software to explain are a number of the query building options and the general approach used in developing a query with I2E. Queries are manually built by combining linguistic items such as words, phrases and sentences, for example. Within these, a user can specify what information the software should be picking up on. In addition to the variety of linguistic items that are used as units with which to build a query, the relationships between items can be edited and adapted in a range of different ways. Prior to detailing the queries built for individual linguistic features applied within this project, it is important to be clear on the tools and materials being used. A list of terms referring to the aforementioned linguistic items and some features of the software used within this project are detailed below.

**Word item**: A unit that contains a single word that is typed in manually. A number of options are associated with the word item. It can be made case sensitive or entered as a substring, for example, meaning that the term can be picked up as a part of a word. For example, the substring ‘psych’ could be entered so that words such as ‘psychiatric’ and ‘psychologist’ would both be picked up. The most important option associated with the word item for this project is the option of including morphological variants in results.

**Morphological variants:** This option makes use of the stemming process mentioned previously. It means that a word will be included in results if it is closely related to the entered word in that it has the same stem or root but varies in the attached morphemes. A morpheme is the smallest grammatical unit in language, a relevant example here would be ‘–s’ at the end of a word to indicate plurality or ‘–ly’ in an adverb. Allowing morphological variants would mean that the entered word and any associated plurals, adverbs, conjugated forms or other variations of the word would be included in results. The option of including morphological variants is indicated to have been allowed for a word when that word is followed by an asterisk.

**Phrase item**: A phrase item allows the user to search for linguistic items appearing together in the text such as two or more words, for example. Within a phrase, the user has the option of requiring whether the words should be in the order entered (ordered) or in any order (unordered) so as to be picked up by the software and constitute a hit, see below for an example. Furthermore, the user can determine how much distance should be allowed between items in a phrase for them to be picked up as a hit. This is measured as ‘word distance’, and the user can enter the maximum number of words that can sit between the items in a phrase in order to be included in the results. These two options allow the query builder to have control over how much variability will be in the results. For example, if a therapist is asking whether their patient has had a good day they might ask ‘Have you had a good day?’ One way of identifying this phrase using I2E would be to create a phrase item ‘good day’. However, this would not pick up a different way of asking that same question such as ‘Has your day been good?’ This problem can be solved by specifying that the phrase item ‘good day’ should be unordered and allow a one word distance between ‘good’ and ‘day’, thus allowing the ‘day been good’ phrase to be included. Both phrases will then be identified by the same query.

**Sentence item**: A sentence item can contain other linguistic items and allows the user to search for the co-occurrence of these items as long as they appear within the same sentence.

**Word class:** A word class is a broad category that can be defined by a given rule (e.g. a verb or a noun) or by an inbuilt or imported dictionary that contains a list of words that qualify for each class or group within the dictionary. For example, the LIWC dictionary was imported into I2E for this project and each category within it, such as negative language or pronouns, would qualify as a word class.

**Region:** A region refers to an element within the structure of the text such as the abstract, introduction or methods sections within an academic paper. In this data set, some regions of interest are ‘date’, provided at the beginning of each session, ‘time’, provided for each message, ‘user’, which identified who sent each message, and ‘text’, the text in each message. In the case of this data set, the documents are in Extensible Markup Language (XML) format, which allows these labels, called tags, to be provided throughout the document. These are part of the coding structure of the data and are not visible to the patient during treatment. Regions of a text are defined and configured at the time of importing a dataset into I2E. A region item then allows the query builder to search for linguistic elements within a specific region (section) of a text. In this data set, this feature is primarily used to include conditions on whether the search is completed within the therapist or patient language as this information is contained within a specific region of the data, called ‘User’. So, to search for a phrase within patient language, the presence of the word ‘patient’ within the ‘user’ region would be a condition for a result.

**Entity: basic:** This refers to any standalone unit of text, this can include numbers, for example. It is often included in a phrase or sentence item when the specific term or unit included is not important to the search or if it is not known.

**Alternatives:** An alternative item allows the software user to create a list of items (words, phrases, sentences, etc.) for which the presence of any of the items in the list will result in a hit (result) for the query. It acts as a series of terms linked by the Boolean operator ‘or’ would in a literature search.

**Negated items:** Items included in a query can be negated. This instructs the software to omit instances where a given, negated, item is present. For example, in an ordered phrase query: ‘a green car’ with one word distance allowed, if the word ‘green’ is negated, the query will pick all other instances of the phrase ‘a ___ car’ with the exception of when the missing word is the word green. In both the software and the figures included below, negation is indicated with the colour red (section 3.3.3 onwards).

**Optional items:** Items can be made optional in a similar way to their being negated. This can be useful in the following type of example. If we consider the following ordered phrase: “Have a good weekend” with no distance allowed between words, the following phrase: “Have a good rest of the weekend” would not be picked up as a hit. One way round this is to include the phrase “rest of the” as an optional item within the larger phrase. In both the software and the figures included below, an optional item is indicated with the colour orange.

#### 3.3.1.1 Iterative query building process

Throughout this project, I2E was used and queries developed with the assistance of the industrial supervisor and experienced members of the Linguamatics staff.

I2E was used to extract features that were built or edited as queries within the software as well as predefined features as I2E allows the importing of external dictionaries. These dictionaries are referred to within the software as ontologies. For each linguistic feature selected for analysis, a query was built within I2E. Further details can be found later on in this section where details of individual queries are provided. Generally, queries were developed following an iterative process of building the query, manually checking results for sources of error, then returning to edit the query before repeating the process (Figure 3-1). This process is repeated until the query builders are satisfied with the output of the query. Improvements in queries can also be verified by comparing sets of results from before and after a change is made.

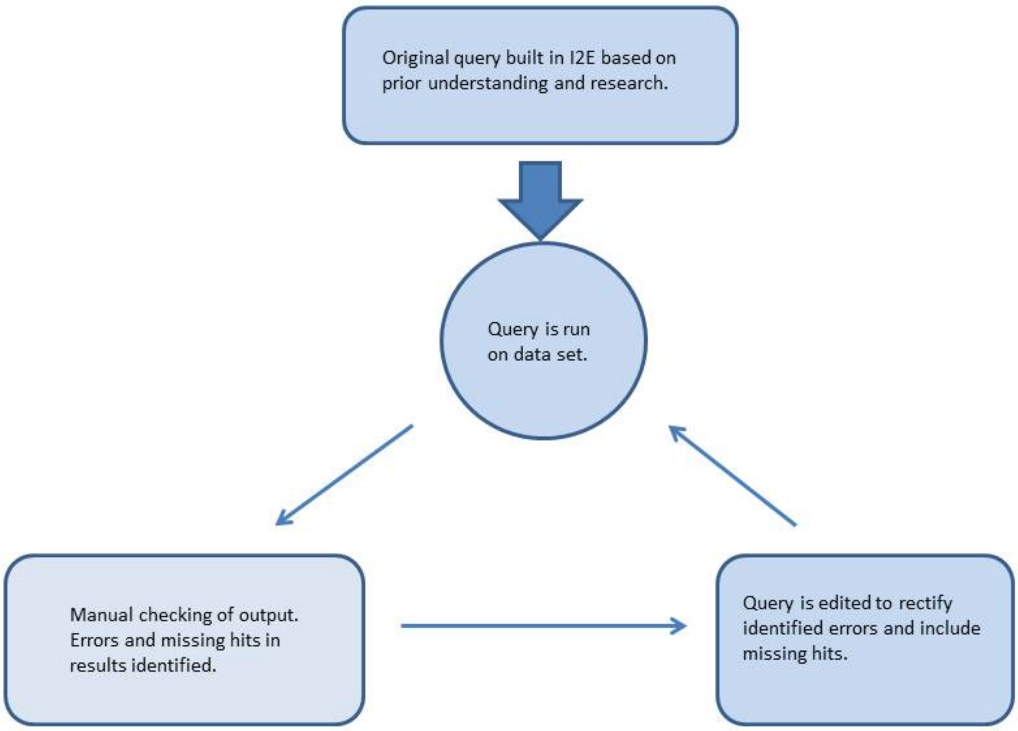


**Figure 3-1 Iterative process of query development in I2E**

Though this iterative process of manually verifying the performance of a query is followed, no specific inter-rater reliability analyses were carried out to support it. This is a limitation of the approach. In future, inter-rater reliability could be performed with raters independently checking query results against their own manual and then comparing their scores.

Following the development of queries, scores for individual features were generated by running the queries on a data set and exporting the emerging results. Results were exported from I2E in the form of a Microsoft Excel spreadsheet that contains frequency counts of linguistic features for each document. In this project, a document refers to a single transcript. Proportional measures of feature use were calculated based on the raw frequency counts for that feature and word counts for each speaker in each transcript and these make up the scores for analysis that are input into STATA. This means that for each appointment, there are patient and therapist scores for each linguistic feature. These are expressed as the percentage of their language that was measured as relating to a given feature. For example, the patient negative language score will represent the percentage of negative language used by the patient over the course of one therapy session.

Four sets of linguistic features were extracted and tested in this project. The origin and development of each of these will be detailed prior to presenting the methods for statistical analysis. I2E was also used to provide a word count for each patient and therapist within each therapy session.

### 3.3.2 Linguistic Inquiry and Word Count features

The Linguistic Inquiry and Word Count (LIWC) is one of a number of word count methods that have been developed and applied within the area of mental health. Originally developed in the context of expressive writing by Pennebaker, Booth and Francis in 2001, it was updated in 2007 and again in 2015 (Pennebaker et al., 2001, 2007; Pennebaker, Boyd, Jordan, & Blackburn, 2015). Popular partially due to its simplicity of use, the LIWC was created through the categorisation of over 3500 terms into a set of approximately 80 categories. These are organised hierarchically, with three overarching categories: Linguistic Processes, Psychological Processes and Personal Concerns. Words within a text are individually recognised as being part of a category and counted as such. A score of frequency of use is then associated with each category of words. It is important to note that categories within the LIWC are not exclusive and a word can fall within one, none or multiple categories.

Given the number of potential categories to investigate within the LIWC, eight categories were selected based on previous work and the literature around depression and anxiety. Higher levels of negative language and first person singular pronouns have repeatedly been found to be associated with diagnoses of depression as well as other mental health disorders such as personality disorders and eating disorders (Arntz et al., 2012; Molendijk et al., 2010; Rude et al., 2004; Wolf et al., 2007). Increasing positive language use has also been associated with improvement over the course of mental health treatment (Arntz et al., 2012). Social orientation, indicated by use of first person plural pronouns and social language has been found to be lower when an individual is not coping well with traumatic or life-changing event (Cohn et al., 2004; D'Andrea et al., 2012; Robbins, Mehl, Smith, & Weihs, 2013). Finally, Insight and Certainty language were selected from the wider group of cognitive mechanisms. Greater use of terms within these categories has been associated with better mental health outcomes (Alvarez-Conrad et al., 2001; Molendijk et al., 2010). Insight was selected based on the focus in CBT on understanding the underlying processes that connect thoughts, emotions and behaviours and certainty was selected as a potential indicator of black and white thinking or openness to change. To recap, the LIWC categories selected were the following: Negative language, Positive language, First person Singular Pronouns (I, me, my, etc.), First Person Plural Pronouns (we, our, etc.), Social language, Insight language, Certainty language, and Negations.

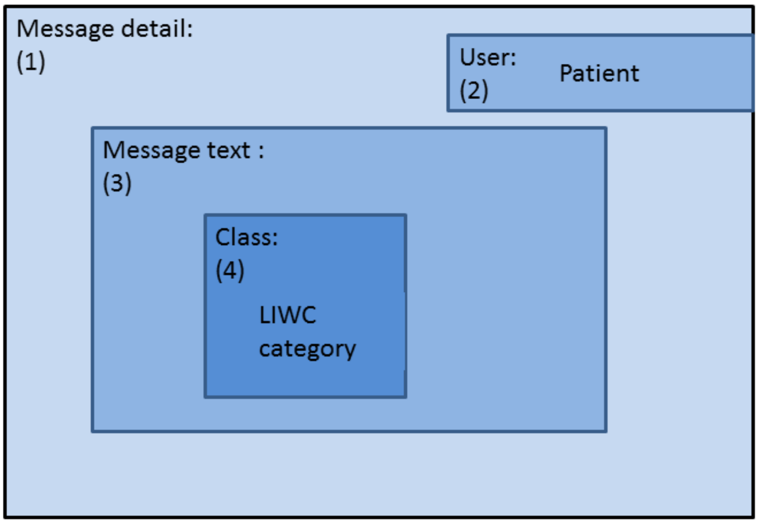


**Figure 3-2 Example LIWC query**

The LIWC dictionary was entered into I2E as an ontology and queries were then built for each LIWC variable selected for this analysis. The queries were designed to pick up and count any term belonging to the relevant category. These were run separately on therapist and patient language so that in the final data set each transcript had a proportional measure of each language variable for both the therapist and the patient. Figure 3-2 provides an example of the structure of these basic LIWC category queries. The outer layer (1), marked as message detail here, indicates to the software that the search is being carried out within the text region called 'Message detail'. Within this there is a region called 'User' (2) which contains information about who is speaking. In this case, the word 'patient' is entered as this example is looking at patient language. The next region (3) is called 'Message text' this refers to the section of the data that contains the actual words typed by the patient. The placing of a word class item (4) within this directs the software to search for instances of a given 'LIWC category' within the message text. Figures will be presented throughout the rest of the chapter to illustrate developed queries or sections of these. For these, only the information contained with it the message detail (3) will be illustrated as the outer structure remains the same for all queries with the only variation being the word 'patient' or 'therapist' in the 'user' (2) region.

### 3.3.3 Sentiment with I2E

I2E text mining queries were developed to build on the sentiment categories from the LIWC (negative and positive language). The aim here was to create a measure that may be more closely representative of sentiment by allowing certain elements of context to be taken into account, primarily the negation of emotional language. The focus here is on measuring what is being written as opposed to how it is being written.

To achieve this, a query was further developed each for positive and negative language. Following the iterative process described above, these were run, in a first instance, as they would be to extract the LIWC measure. When a query picks up an element in the text, this is called a ‘hit’. The hits (results) from a query are read through to check whether what is being picked up in the transcripts is concordant with what the query intends to pick up. This relies on human opinion so the process is inherently subjective.

The two sentiment queries were developed as follows.

#### 3.3.3.1 Negative language query

The query was created as a set of alternatives, containing the three phrases that will be detailed in the next paragraphs.

1) Excluding phrases apologising for poor Internet or device connection or function.

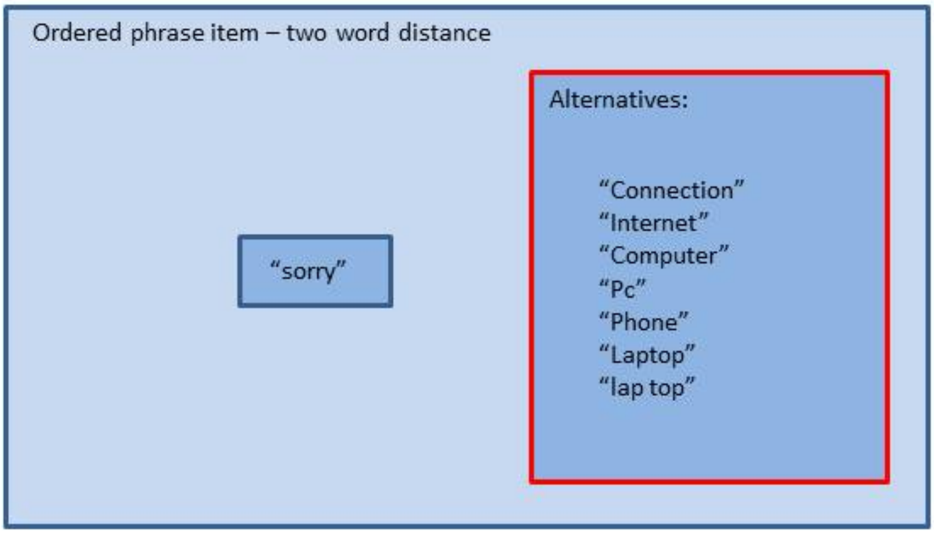


**Figure 3-3 Negated phrase relating to technical issues**

The word 'sorry' was very common in the development data set that was worked with and this was often in the context of the patient or therapist expressing an apology for a technical issue. It was decided that this may detract from focusing on negative sentiment coming from the individual typing and would therefore be excluded from counts of negative language. This was done by creating a phrase item that would include the word 'sorry' followed by one of a number of possible words or phrase listed in the set of alternatives in Figure 3-3.The phrase was set as ordered and allowed a distance between items of up to two words. This was to allow for the multiple ways an individual could associate the two parts of the phrase. These phrases were excluded from results of the negative language query by instructing I2E to omit them.

2) Excluding negated negative language whilst including negative language alone.

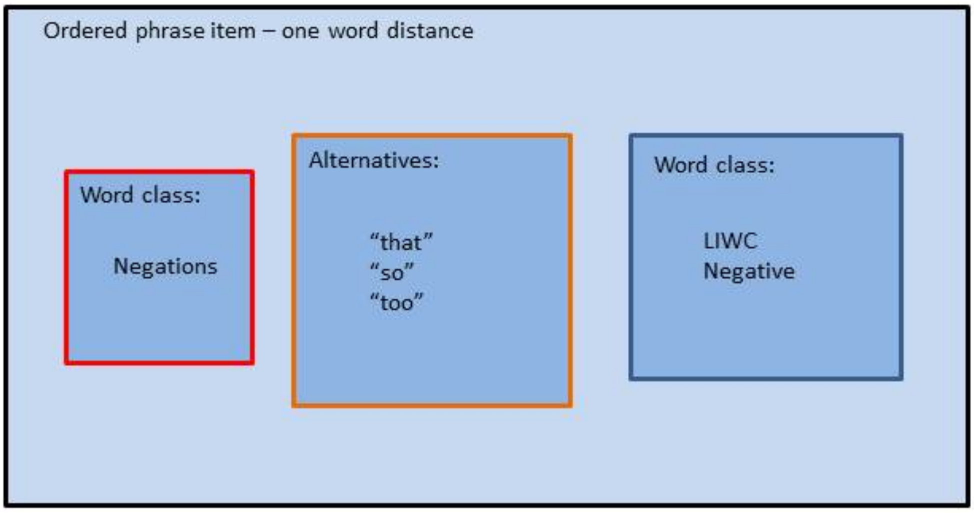


**Figure 3-4 Including negative affect**

This phrase was built around the negative affect word class that formed the original LIWC query. The negative affect class was entered into a phrase and the ‘negations’ class preceded it, with a list of alternatives between the two that was made optional. The ‘negations’ class was itself negated, meaning that the query would only pick up instances where a negative affect word was not preceded by a negation, as this would change the valence of the phrase. With the query built this way, the phrase ‘bad’ was picked up as a hit for negative affect, whereas the phrase ‘not bad’ wasn’t. A one word distance was allowed between items in this phrase. The addition of a set of alternatives as an optional item meant that if the listed common qualifiers were included in the phrase as well, it would still be picked up. For example, the common phrase ‘not actually that bad’ would be picked up as a negative language hit without the additional list of optional terms as ‘not’ is two words away from ‘bad’ and thus t0o far for the negation to exclude the hit. The aim was to exclude this type of phrase as it was decided that it did not qualify as negative language. The inclusion of the optional set of alternatives provided a method to solve this problem.

3) Including negated positive affect

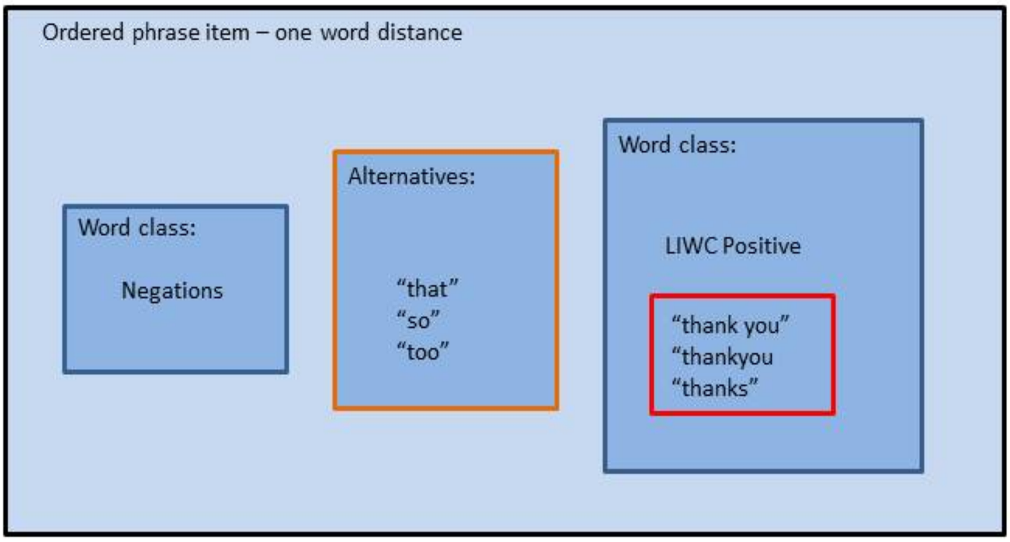


**Figure 3-5 Including negated positive affect**

This phrase was developed to achieve the opposite of the previously described phrase; to include negated instances of positive language. This was built by creating a phrase item containing a 'negations' class placed in front of the positive language class item. Within the positive language item, a list of three alternatives is negated. These are 'thanks', 'thankyou' and 'thank you'. This is due to phrases such as 'No, thank you' that would be picked up as negative language when this is not what is expressed. On the other hand, phrases such as 'wasn't happy' would be picked up by the query and included in results for negative language.

#### 3.3.3.2 Positive language query

The positive language query was built as a multi-query, meaning that two or more queries were combined. In this case it can be seen as a simple subtraction operation with the results from query 2 being subtracted from the results of query 1.

#### 3.3.3.2.1 Query 1

1) Exclude negated positive affect and include positive affect

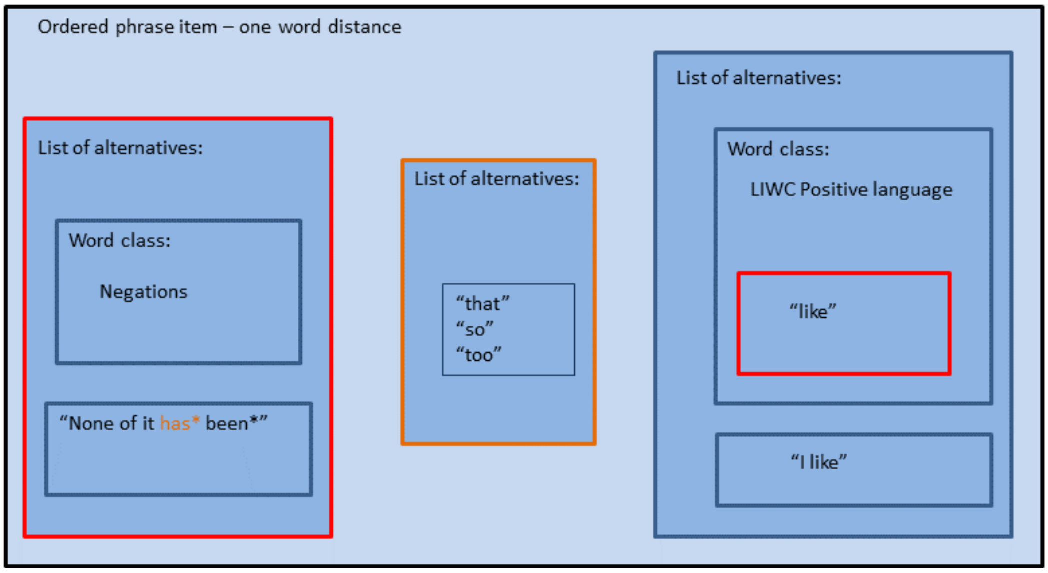


**Figure 3-6 Include positive affect, exclude when negated**

This phrase, illustrated in Figure 3-6, aims to pick up evidence of positive language but not where this is preceded by a negation ('not'), making it negative. This section of the query also mirrors that in Figure 3-4 that was described above, with a couple of extra elements. The phrase can be subdivided into two items; a smaller phrase (left in image) and a list of alternatives (right in image). The smaller phrase contains two parts. The first was negated, meaning that results that contain these elements should be excluded from the results. It was made up of a negations class and a phrase: 'none of it has been' in which morphological variants were allowed for the words 'has' and 'be' and 'has' was an optional word (indicated here by the colour). Morphological variants were allowed for the words 'has' and 'been' so that variations such as 'had been' were also negated. The phrase 'none of it has been' followed by a positive word was added to the negated list of alternatives as it was a common source of negative affect in the development set and it was determined that these instances should be excluded when looking to pick up positive language. This phrase also contains the optional alternative list of qualifying words ('that', 'so', 'too') used and described previously. The second alternative list making up the phrase is based on positive affect language with the exclusion of the word ‘like’ alone, but the inclusion of the phrase ‘I like’. This change was made give the frequency of the use of the word ‘like’ as a filler or non-affective term.

2) Include negated negative affect

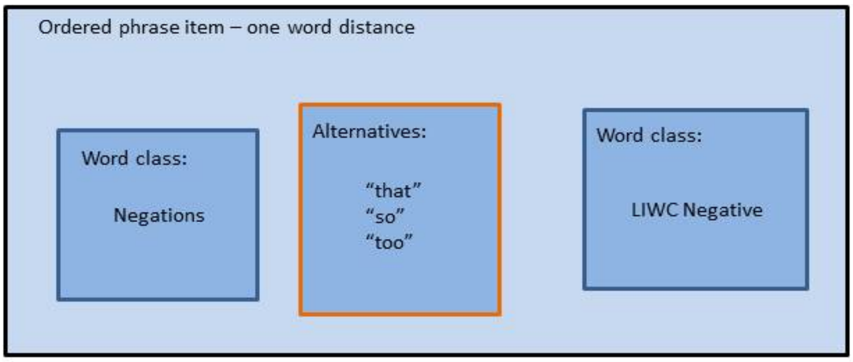


**Figure 3-7 Include negated negative affect**

This was built similarly as was described within the negative language query above (Figure 3-5) but aimed to achieve the opposite. A negation class was included in the phrase in front of negative language and so that phrases such as ‘not too bad’ would be included in the hits. The optional list of alternatives was also included here for the same reasons as described above.

#### 3.3.3.2.2 Query 2

These are the elements that were subtracted from the results of Query 1. This means that though they may have been picked up as results through query 1, they were then removed from the results and not counted towards the positive language score.

1) Social conventions

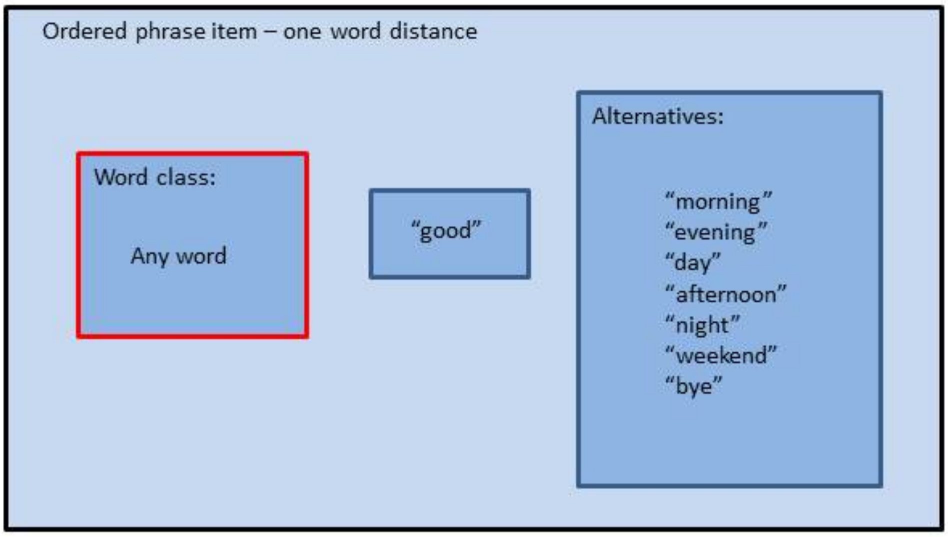


**Figure 3-8 Social conventions**

This was a phrase built to remove the use of conventional social phrases from positive language counts. Most, if not all, transcripts begin with a phrase such as 'good morning' or 'good evening' from either the patient, the therapist or both. This was seen as an element that was unnecessary given that it would not differentiate individuals if it were present across transcripts.

The phrase combined the word 'good' with a list of alternative terms that could be used as a greeting. A negated word item also preceded the term 'good' so as to only exclude these terms when they were used at the beginning of an utterance. The aim here was to avoid removing a phrase such as 'I had a good day' from the count of positive language.

2) Common neutral and filler terms

This phrase aimed to remove words such as 'ok' and 'well' from being counted as positive affect when they were used as filler terms. The phrase is made up of two sets of alternatives, the first being optional. A negated word item was used to ensure that this query would affect only these terms when they were at the beginning of an utterance, where they were most likely to be used in a non-affective context. The optional alternatives 'ok' and 'as' allowed for phrases such as ‘as well’ and ok well’ to be removed from the positive language counts.

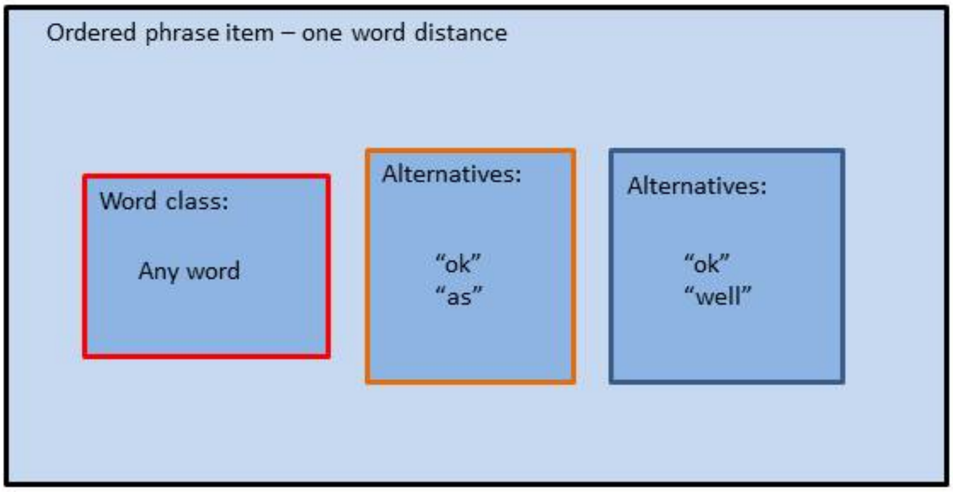


**Figure 3-9 Neutral and filler terms**

### 3.3.4 Positive and Negative Affect Schedule

#### 3.3.4.1 Background

The Positive and Negative Affect Schedule was developed as a self-report scale to assess affect by asking individuals to rate how much they are or have been feeling a particular emotion on a scale of one to five. In the original version, developed by Watson and Clarke, there were only twenty terms, ten for each of positive and negative affect (Watson, Clark, & Tellegen, 1988). This included words such as ‘irritable’, ‘hostile’ or ‘distressed’ for negative affect and ‘inspired’, ‘enthusiastic’ and ‘alert’ for positive affect. A revised and expanded version called the ‘PANAS-X’ was published in 1994 and the manual updated in 1999 (Watson & Clark, 1999). It contains 60 terms that were split into eleven categories. Seven of these categories can also be considered subcategories of negative and positive affect. These are: Anger, Hostility, Guilt, Sadness, Joviality, Self-Assurance and Attentiveness. As the focus here is on affect, these are the categories of interest in this part of the project. There are four further categories: Shyness, Fatigue, Serenity and Surprise that will not be included in the work in this project. The PANAS-X has been used in a range of research work including looking at the effects of positive affect on broadening attention and improving coping skills (Fredrickson & Joiner, 2002) or in work looking at affective features of borderline personality disorder (Trull, Useda, Conforti, & Doan, 1997). Evidence for validity and reliability were put forward in the PANAS-X manual (Watson & Clark, 1999).

#### 3.3.4.2 Expanding the PANAS-X

For this project, I considered that the PANAS-X scale may be another option from which to build linguistic measures to apply to the therapy transcripts. It contains a narrow set of terms that are subdivided into different subcategories than the LIWC. Within negative language, there are four subgroups and within positive language there are three subgroups. However, the PANAS-X was developed as a self-report rating scale and is made up of only 60 words, making it a limited dictionary to use for linguistic analysis. Due to this, I2E was used to expand the dictionary by using the transcripts in the development data set to 'harvest' other words used by patients that represented the same emotions and feelings included in the PANAS-X. This also adapted the dictionary to the context within which the project was being carried out. Thus, this was an exploratory and experimental approach to this task.

##### 3.3.4.2.1 Harvesting relevant terms

To harvest these words, I2E was used in a three stage process that involved 1) determining the linguistic contexts of the terms in the PANAS-X dictionary, 2) searching for other words within these contexts and 3) manually verifying and including the harvested terms and phrases into an expanded PANAS-X dictionary.

1) The first step involved building a query with I2E that was made up of a phrase that contained three items. The central item was a PANAS-X category and the items either side were word class items that would pick up any word used before or after the PANAS-X category. This phrase would

therefore provide, in the results, the words most frequently found surrounding the terms in a given PANAS-X category.

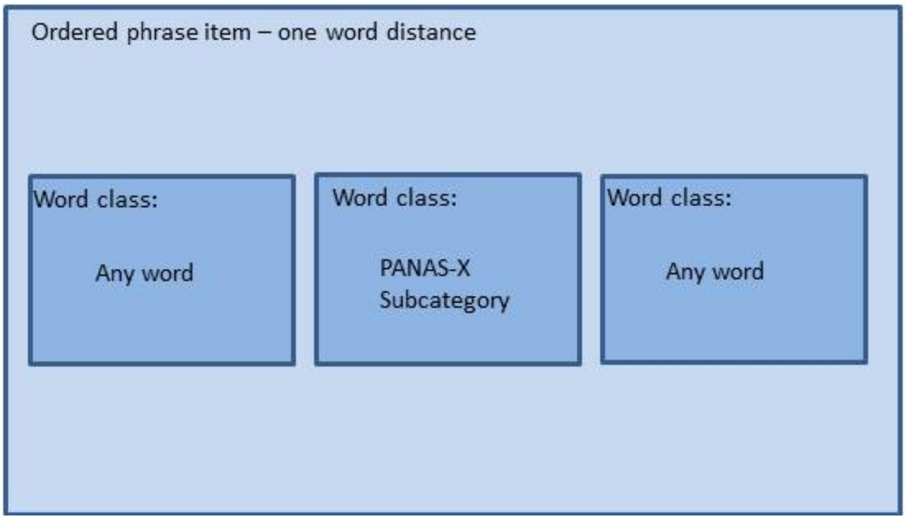


**Figure 3-10 Search 1 in PANAS-X expansion**

The structure of this first search is illustrated in Figure 3-10. If this was carried out with the PANAS-X subcategory ‘Hostility’, a result of the search could be the phrase “I get very frustrated with”

2) The second stage of this process involved building a query from the results of step 1. That is to say to create a phrase item containing the frequently used terms found either side of the PANAS-X category and placing a word class item within the phrase looking to pick up any word used. This can be considered as a blank space to fill. The results of this query were then considered potential candidates with which to expand the dictionary.

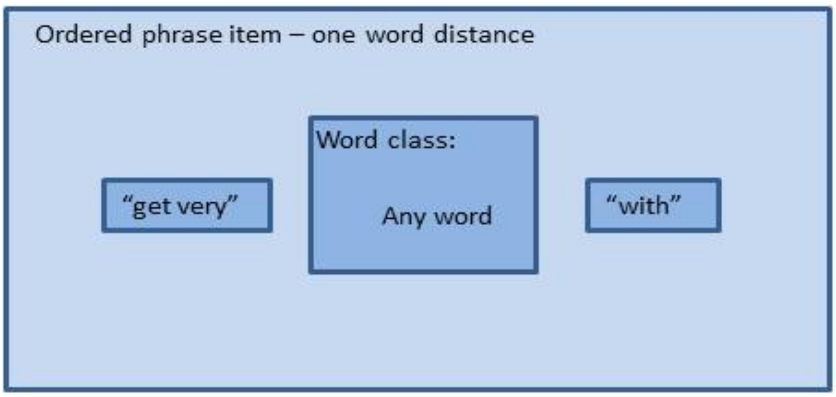


**Figure 3-11 Example of search 2 in PANAS-X expansion.**

An example of the type of query that might be used in this second process is illustrated in Figure 3-11. This uses the results from the first search, in this case using the example of ‘get very frustrated with’.  The results to this search could be phrases such as: “get very upset with...”, “get very angry with...” The terms ‘angry’ and ‘upset’ would then be potential candidates with which to expand the PANAS-X subcategory ‘Hostility’.

However, other phrases such as ‘get very carried away with’ that are not synonymous with ‘frustrated’ were also likely to be picked up by the software. This means that the results then needed to be filtered by hand to check whether they were relevant to the category of emotion being considered.

3) The final stage of this process involved manually checking the candidate terms put forward by the harvesting process described above and making a judgment about whether these fit into the categories they were selected for or should be discarded as errors.

**3.3.4.2.2 Word2Vec**

Additionally to this, I looked to expand the dictionary using more objective methods. There is a method called ‘word2vec’ in which words within a text corpus are represented by vectors. These vectors are calculated based on the patterns of co-occurrence of those words. A number of different computer codes have been developed to carry out this operation. Within this project, an implementation of word2vec developed by Google was used. This means that the computer code used to perform the relevant operations is provided by Google. The vectors associated with each word were sourced from the default dictionary associated with the code and individual terms from the PANAS-X dictionary were then run through the code in order to extract the 10 other terms that were most closely related to them statistically. For each word in the PANAS-X dictionary, the 10 words with the closest vectors were returned. These are expected to be the closest neighbours of the input words and word2vec therefore allows objective expansion of the PANAS-X dictionary. These terms were also manually verified as the system can return antonyms which I was looking to exclude.

#### 3.3.4.3 Creation of a new dictionary

Once the process above had been repeated for each of the PANAS-X categories, a new dictionary was created that includes the emotion and feeling words from the original categories as well as those harvested using I2E and Word2Vec that fit the categories when checked manually. Phrases were also included, such as 'lose my temper' in the hostility subcategory, for example. Each category would then count as a linguistic variable to be tested at a later stage. The expanded version contained 383 terms over the seven included subcategories: Hostility (67), Guilt (22), Sadness (74), Fear (40), Joviality (80), Attentiveness (28) and Self-Assurance (72).

#### 3.3.4.4 Sentiment queries

As with the LIWC categories of negative and positive language, queries were developed for negative and positive language based around the expanded PANAS-X categories.

##### 3.3.4.4.1 Negative language with Expanded PANAS-X

Similarly to the LIWC-based query for negative language, the PANAS-X based negative language query was made up of a set of three alternative phrases.

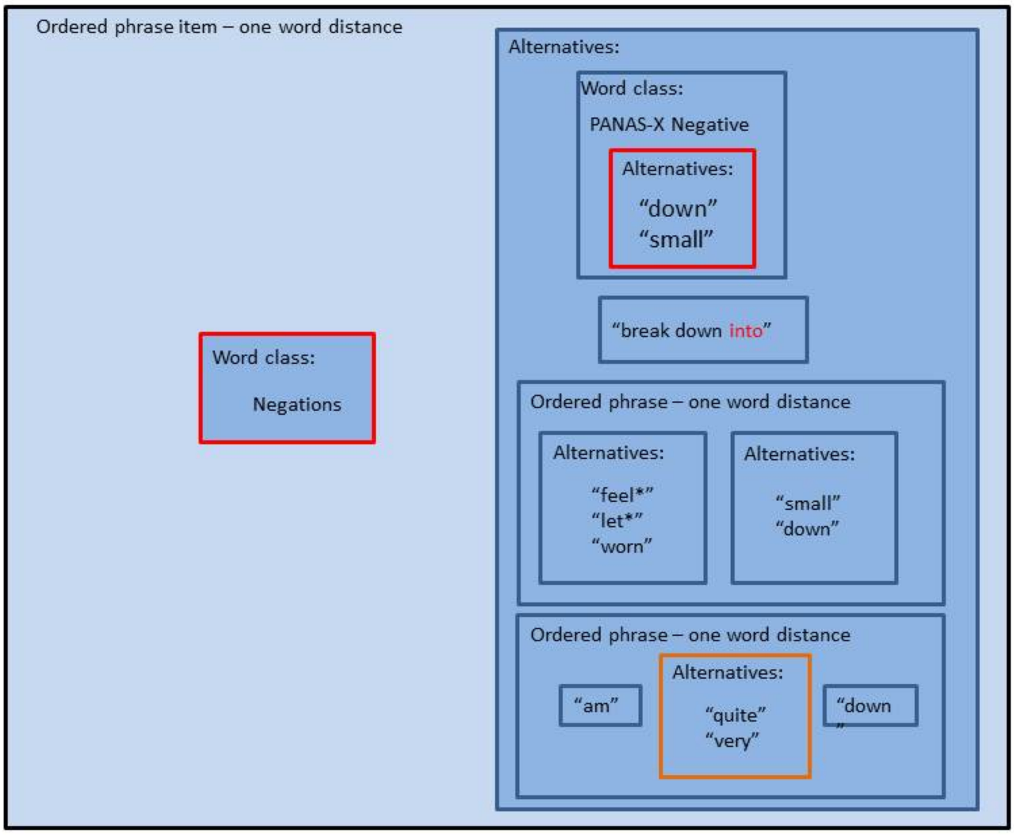


**Figure 3-12 Negative language in PANAS-X**

Negated negative language was excluded whilst looking to pick up terms and phrases that were determined to qualify as negative language. This query differs from the LIWC query as it takes account of the presence of the words 'down' and 'small' in the expanded PANAS-X negative language category. The terms 'small' and 'down' were found to be commonly associated with negative affect in the development set but primarily in the context of qualifying words such as 'feel' or 'break' whereas alone, the terms were used in a variety of different contexts that mostly did not qualify as expressing negative affect. To avoid the query picking up irrelevant terms and phrases, the two ambiguous words were removed from the Negative affect class by negating the individual terms and entered into the query separately within the context of relevant phrases such as: 'break down', 'feel small/down' or '[I] am down'.

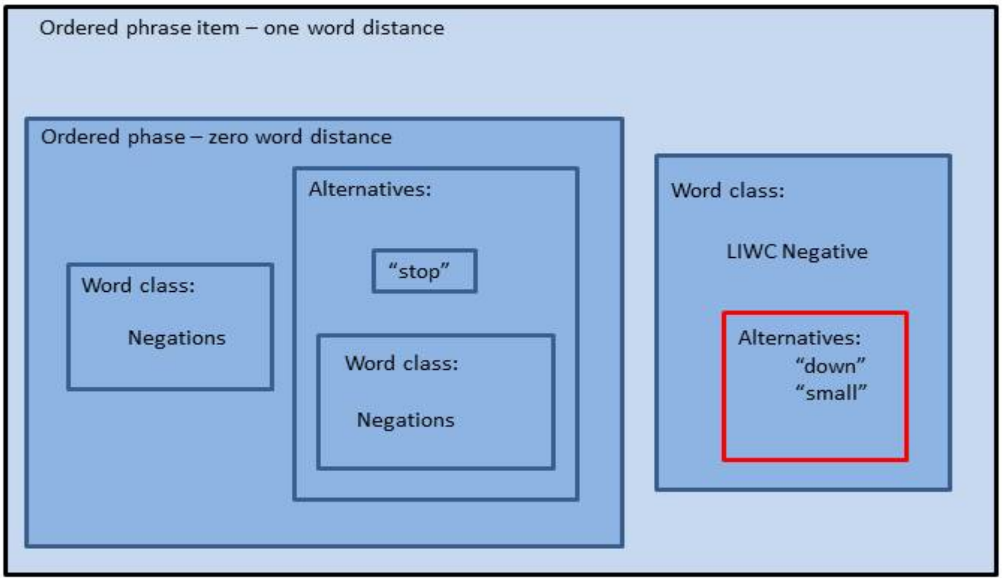


**Figure 3-13 Include double negated negative language**

The phrase illustrated in Figure 3-13 was added to the set of alternatives as a phrase to be picked up by the software to remedy the exclusion of instances where a negative language term was preceded by two negations. This creates a double negative such as 'can't not' and maintains the valence of the negative term that follows. The word 'stop' was also added as an alternative to a second negation as this was a common occurrence in this data set. The inclusion of this item therefore means that phrases such as 'can't stop worrying' or 'can't not worry' will be picked up as negative language by the query.

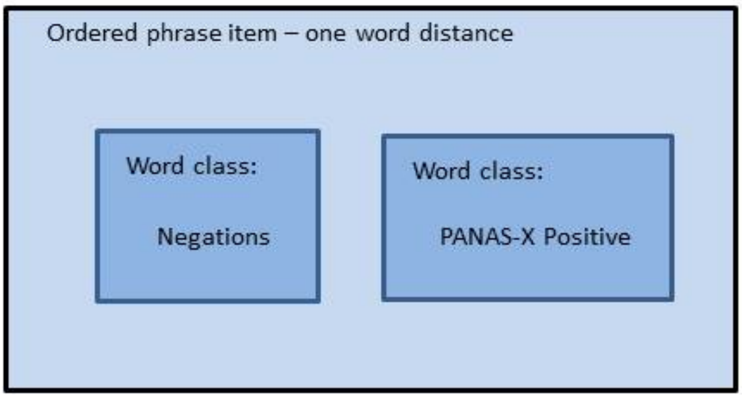


**Figure 3-14 Include negated positive language**

The phrase illustrated in Figure 3-14 is a simple phrase that included phrases made up of a positive language term preceded by a negation in the results for the overall negative language query. Phrases such as ‘not very excited’ would be picked up by this query. This is a slightly simpler version of the phrase in Figure 3-5.

#### 3.3.4.4.2 Positive language with Expanded PANAS-X

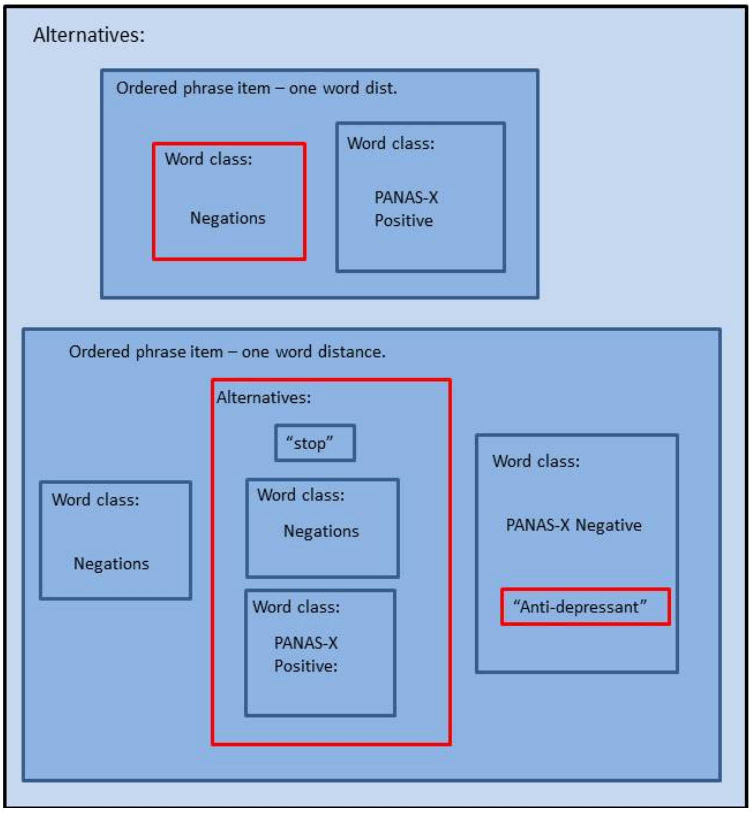


**Figure 3-15 Query 1 PANAS-X positive language**

The positive language query based on the expanded PANAS-X dictionary was built in a similar way as the LIWC-based positive language query with the results from one query being subtracted from the results of another. It is, however, a slightly simpler version.

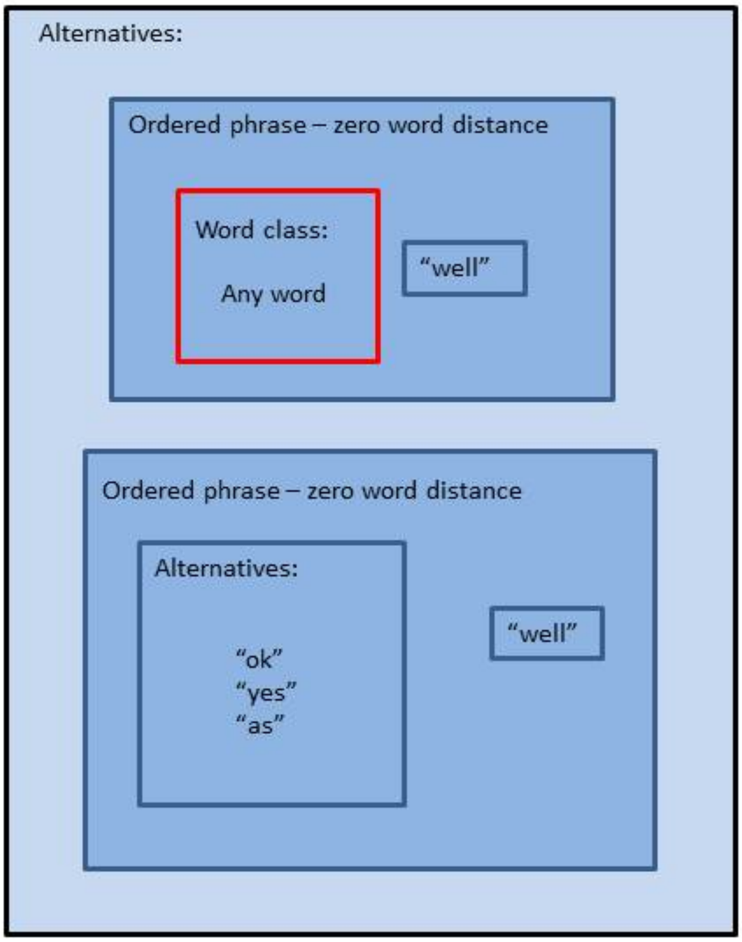


**Figure 3-16 Filler words in positive PANAS-X query**

In the first query, illustrated here in Figure 3-15, negated positive language was excluded from the results and negated negative language was included while positive language alone was also included. The additional negated set of alternatives plays the same role as that in Figure 3-13, to avoid including negative affect terms preceded by a double negative. The second query, illustrated in Figure 3-16 is made up of two phrases looking to pick up the use of 'well' as a filler or non-affective term so as to remove these hits from the results of Query 1. The first phrase picks up 'well' when it is at the beginning of an utterance, this is indicated by the negation of the 'Any word' word class. The second phrase picks up the phrases 'yes, well', 'as well', and 'ok, well'. Phrases around the word 'like' were not included here (as they had been in Figure 3-9) as the word 'like' was not included in the PANAS-X positive terms dictionary. These results are subtracted from the results of query one and therefore did not count towards the count of PANAS-X positive language.

### 3.3.5 Revised Cognitive Therapy Scale

#### 3.3.5.1 Background

The Revised Cognitive Therapy Scale (CTS-R) (Blackburn et al., 2001) is a scale that was based on the original Cognitive Therapy Scale (Young & Beck, 1980) and provides a framework for rating Cognitive Behaviour therapists. The scale is based on the concept of the cognitive cycle. This revolves around conceptualisation and involves the interconnections of thoughts, feelings and behaviour. The idea is that changing these four elements will lead to changes in a patient's conceptualisation of their experiences and that this is the core process important in improving their mental state. Cognitive behaviour therapy is a very structured form of therapy and the CTS-R aims to evaluate both how closely therapists are keeping to the framework and their skill in doing so. There are twelve CTS-R items that focus on different aspects of the conceptualisation cycle as well as the skills a therapist needs to encourage a patient to move between these and make changes. The scale is separated into distinct elements such as agenda setting, collaboration, pacing or feedback. In the 2001 manual, the twelve items include five general items and eight cognitive therapy specific items. The agenda setting and adherence items are considered both general and specific and so are included in both categories. The CTS-R is often used as an assessment tool in the training or ongoing practice of CBT therapists (Keen & Freeston, 2008).

#### 3.3.5.2 Selection of items

Originally, the aim in this part of the project was to develop a text mining query with I2E for each item of the CTS-R. However, given the exploratory nature of this project, the time demands associated with developing complex queries and the subjective nature of some of the items on the CTS-R, it was decided that a smaller subset of items would be the focus of query development within this project. These items were selected based on the literature concerning therapeutic elements within cognitive behaviour therapy, the practicality of converting them to text mining queries and discussion with clinical staff at Ieso Digital Health. In terms of the practicality of transferring items from a manual scale to an automatic version, the difficulty lies in items that rely on judging appropriateness of therapist behaviour in any particular situation and for any particular patient. For example, three items involve rating the therapist's ability to elicit appropriate emotion, cognitions and behaviour. This means that to provide a score for these items, the rater needs to consider whether emotions, cognitions or behaviours are being brought into the conversation and discussed in the treatment session but also whether these are appropriate for a particular patient and whether this has been done with the right level of skill, making these items difficult to automate at this stage of the research.

There has been some previous research looking at aspects of cognitive behaviour therapy and their association with therapy outcome. A recurring idea within this research has been that of the presence of two factors within therapist rating scales. One factor relates to structural items or adherence to therapy protocol and the other to therapist skill or competence (Brown et al., 2013; Whisman, 1993). The structural items appear to have the largest objective aspect to them and thus were considered the best candidates for automation at this point. After consideration of practicality for transfer to automatic methods and clinical relevance, four items within the CTS-R were chosen. These were Agenda setting, Homework setting, Pacing and Interpersonal Effectiveness. The first three items were determined to be most practical for query development in that they rely less on personal judgment than some of the other items while being very important to adherence to the CBT framework and the therapy outcome. These are the items that tend to make up the 'structure' factor of therapist rating scales (Whisman, 1993). Interpersonal Effectiveness is also often referred to as an indication of therapeutic alliance and is seen to have a great deal of influence on clinical outcomes (Lambert & Barley, 2001; Martin, Garske, & Katherine, 2000). Although it was likely to be a more difficult task in terms of automation as a text mining query, the suggested clinical value of the item is so great that it was decided that this should be attempted.

#### 3.3.5.3 Agenda setting

According to the CBT framework, therapists are expected to set an agenda with patients at the beginning of each treatment session. This is one element of the structured nature of CBT. In the CTS-R, there is also a subjective element to this item and the score allocated is meant to reflect both the presence of an agenda as well as the appropriateness of the items included. At this stage, however, the focus for query development was on determining the presence or absence of an agenda within a therapy session only, not the value of the items within it. The I2E query developed therefore aimed to pick up on therapist language that suggested that an agenda had or was being set.

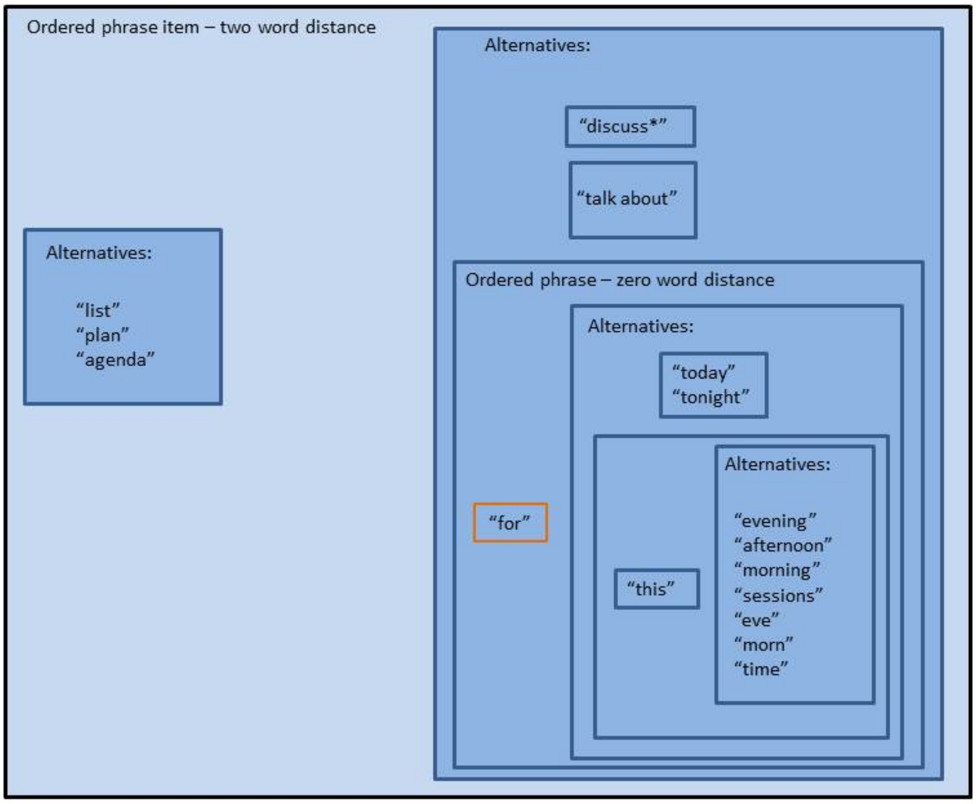


**Figure 3-17 Agenda setting query 1**

As with a number of the previously described queries, this query was made up of a set of three alternative items, all aiming to pick up different ways that a therapist might suggest or refer to setting an agenda. Therapists did not always refer explicitly to an agenda but may, for example, have asked what the patient would like to discuss in their session. This meant that a number of variations of this kind of phrasing were included in the query.

The phrase item illustrated in Figure 3-17 aims to bring together a first set of phrases that a therapist may use to refer to setting an agenda. It was made up of two sets of items. The first included different terms a therapist may use for ‘agenda’ and the second, the terms that suggested the setting of an agenda or plan for the session. This query aimed to pick up phrases like ‘a list for this session’ or ‘the plan today’.

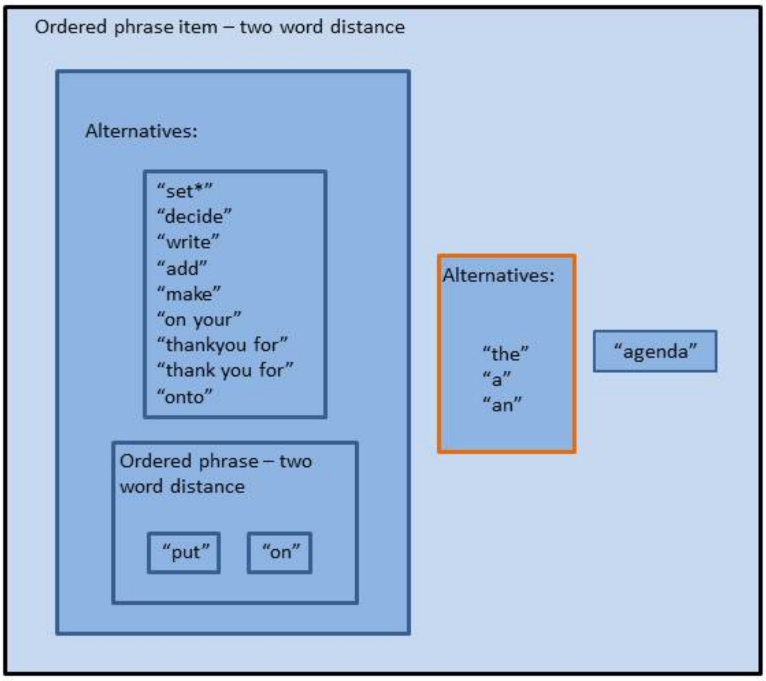


**Figure 3-18 Agenda setting 2**

The second phrase in the set of alternatives making up the agenda setting query is illustrated in Figure 3-18. It aimed to pick up another set of phrases that may have referred either to the setting of an agenda at the beginning of a session (e.g. ‘write the agenda’) or to an agenda that had previously been set (e.g. ‘thank you for sending me an agenda’).

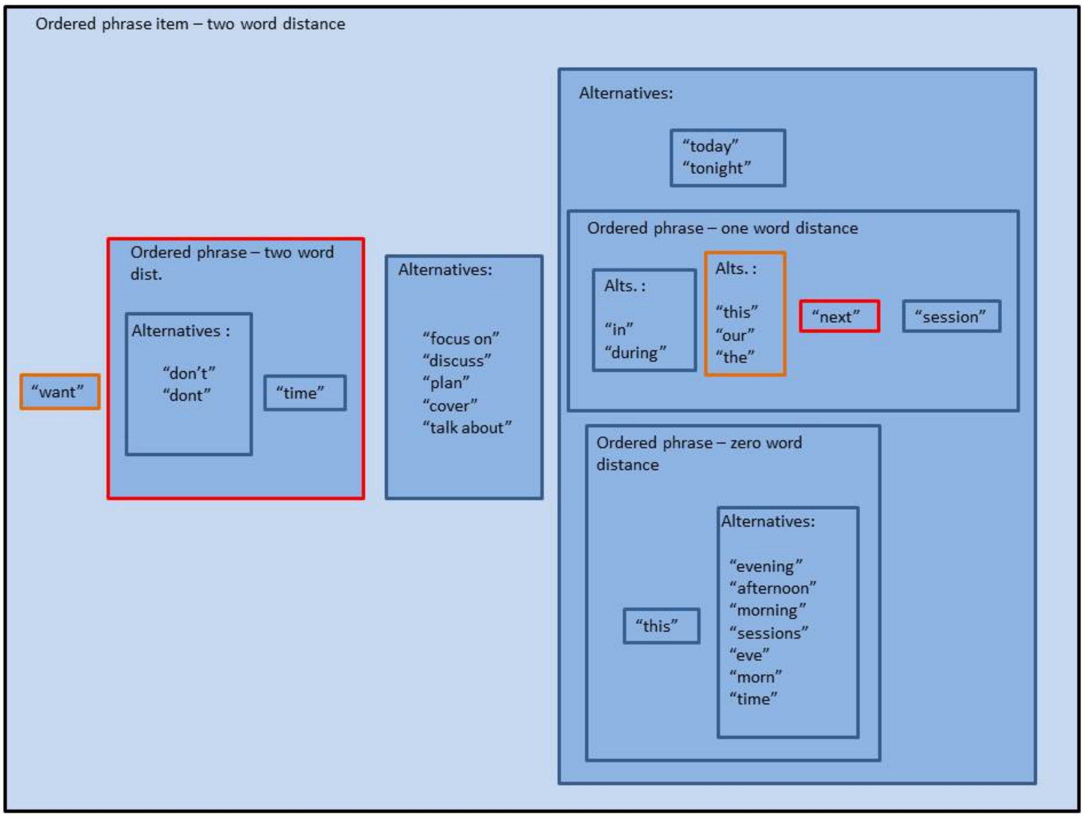


**Figure 3-19 Agenda setting query 3**

The final alternative phrase included in the agenda setting query is illustrated in Figure 3-19. This was a slightly more complex phrase that aimed to pick up a therapist's reference to setting an agenda when the word 'agenda' had not necessarily been used. The center of the phrase was a list of alternative verbs that a therapist may have used when planning what the session would cover (e.g 'focus on', 'discuss'). The following set of alternatives (on the right) listed a number of time qualifiers that would disambiguate agenda setting type phrases from other uses of verbs such as 'focus', 'cover' or 'discuss'. Within this set of alternatives, the second alternative from the top aimed to pick up phrases such as 'in this session', whilst excluding references to the next session as these were determined to not be relevant when the scoring was related to the particular session being analysed. The inclusion of the negated phrase in second position of the query removed from the results the phrases where a therapist might have said that they 'don't have time to focus on' something 'this session', which would count against evidence of a well-set agenda.

#### 3.3.5.4 Homework setting

A second element that is very important within CBT is homework. This refers to tasks for the patient to complete between sessions in order to either learn more about certain CBT concepts and understand the process, or practice skills or techniques discussed with a therapist. Appropriate homework should be decided on through discussion between the therapist and patient. Similarly to Agenda Setting, the CTS-R item of Homework Setting includes a subjective element of rating the appropriateness of allocated homework tasks. Again, at this stage of research, the focus in query development was solely on determining whether there was evidence, in the language used, of homework having been set.

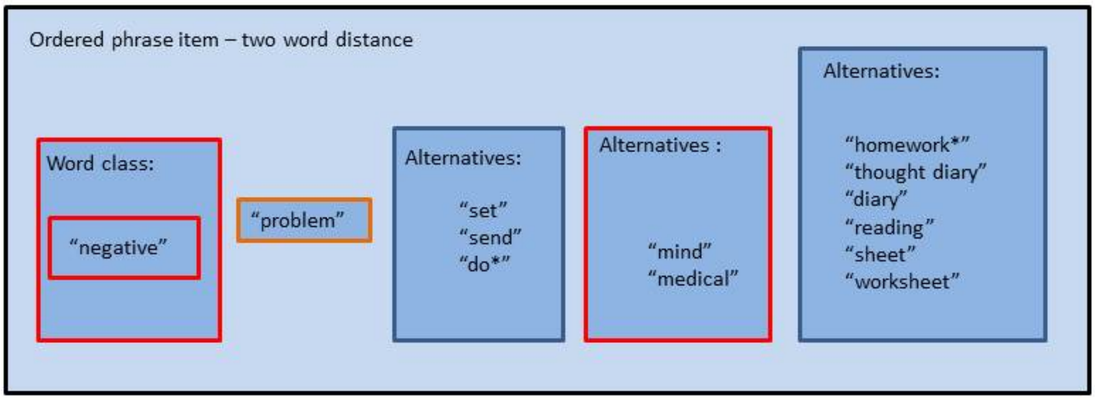


**Figure 3-20 Homework Setting**

The query, illustrated in Figure 3-20 aimed to pick up evidence that a therapist had discussed and set homework with their patient. The query was an ordered phrase made up of a number of items, primarily alternatives. The two sets of alternatives in blue, and therefore included in the query, were a set of verbs that were found to be associated with homework setting and a set of terms used to express the idea of 'homework' or various common homework tasks. The set of potential homework tasks was developed based on clinical knowledge, therapist training materials and manual reading of a set of transcripts from the IPCRESS clinical trial completed in 2009 (D. Kessler et al., 2009).

#### 3.3.5.5 Pacing

The pacing item in the CTS-R refers to the therapist's ability to maintain the timings and pacing within a session so as to ensure that all agenda items were covered by the end of the session. Sessions were either 30 or 60 minutes long and could go very quickly, so ensuring the session is on track in terms of timing is very important. Once again, there is an element of subjective judgment involved in scoring the pacing item associated with the appropriateness of a therapist's pacing of a therapy session.

The development of the query for the pacing item primarily revolved around picking up phrases that contained a time element and then determining whether these were referring to time within the session, which would be relevant, or in relation to something external. Another step of query development was based on phrases that a therapist might have used in order to attempt to move the session forward. These are phrases such as 'let's move on' or 'the next item'. The query for the pacing item was made up of a large number of alternatives. For practical reasons, the illustration of this query has been split into two parts, but the full query puts all alternative items within both of these into one larger set of alternatives. Part 1 of the query is illustrated by Figure 3-21.

The first three items are phrases a therapist might use to make explicit the fact that they are aware of how much time is left in the session. The final three phrases in the first part of this query are variations on ways a therapist may look to move the session along such as references to the 'next item' on the agenda or encouraging the patient to focus on the next task by using phrases such as 'let's move on to'. In the penultimate phrase (vertically), a set of terms was negated in front of 'moving on to' so that phrases including those terms would be omitted from the results. This was done because without this there would be some overlap between the results for this phrase and the next phrase listed below. Without this condition, the phrase 'let's move on to' would have been picked up twice because it would qualify within the 'moving on to' phrase and the 'let's move on phrase'. As the query is defined here, however, it is only picked up through the last phrase as ‘let’s move on’.

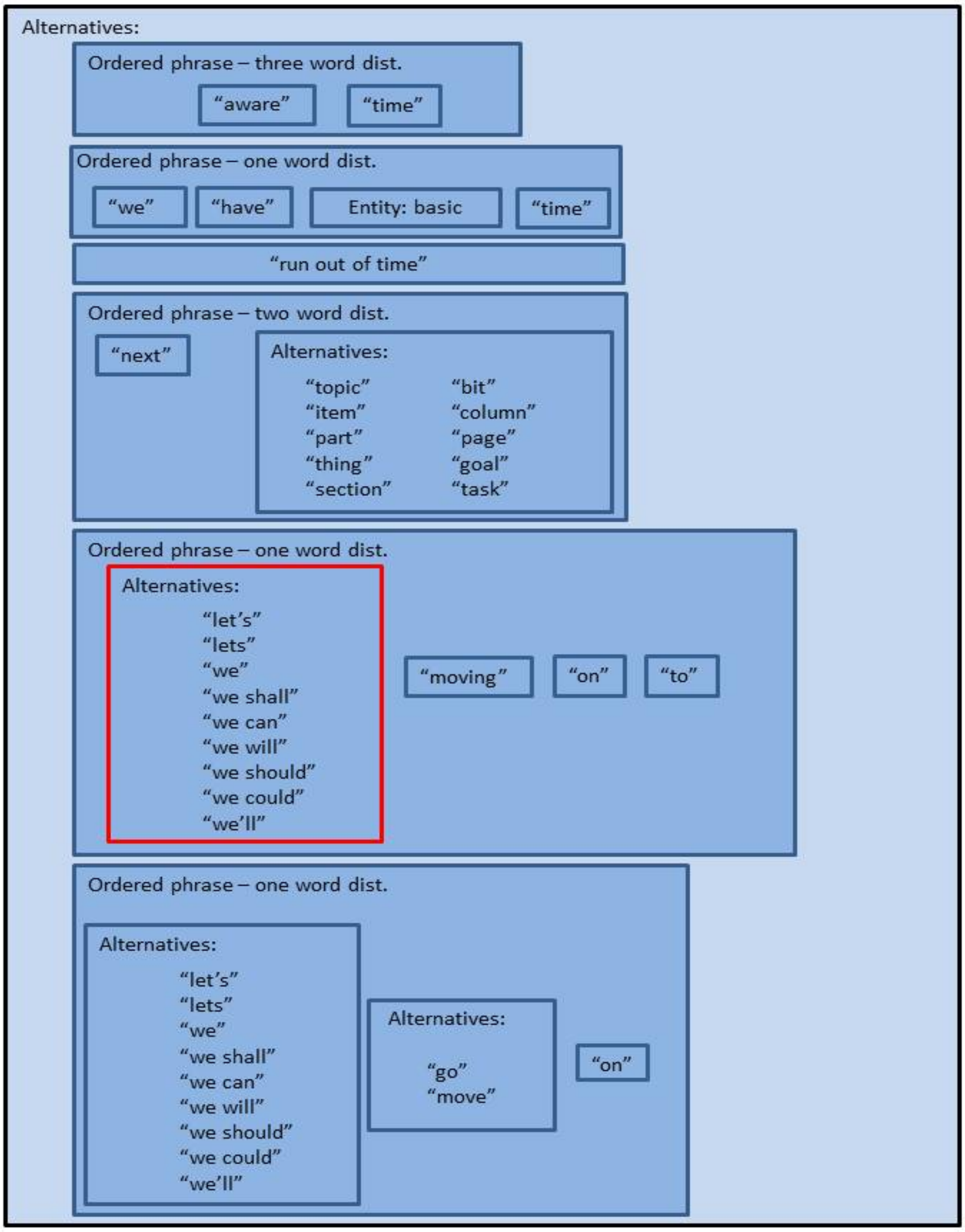


**Figure 3-21 Pacing query part 1**

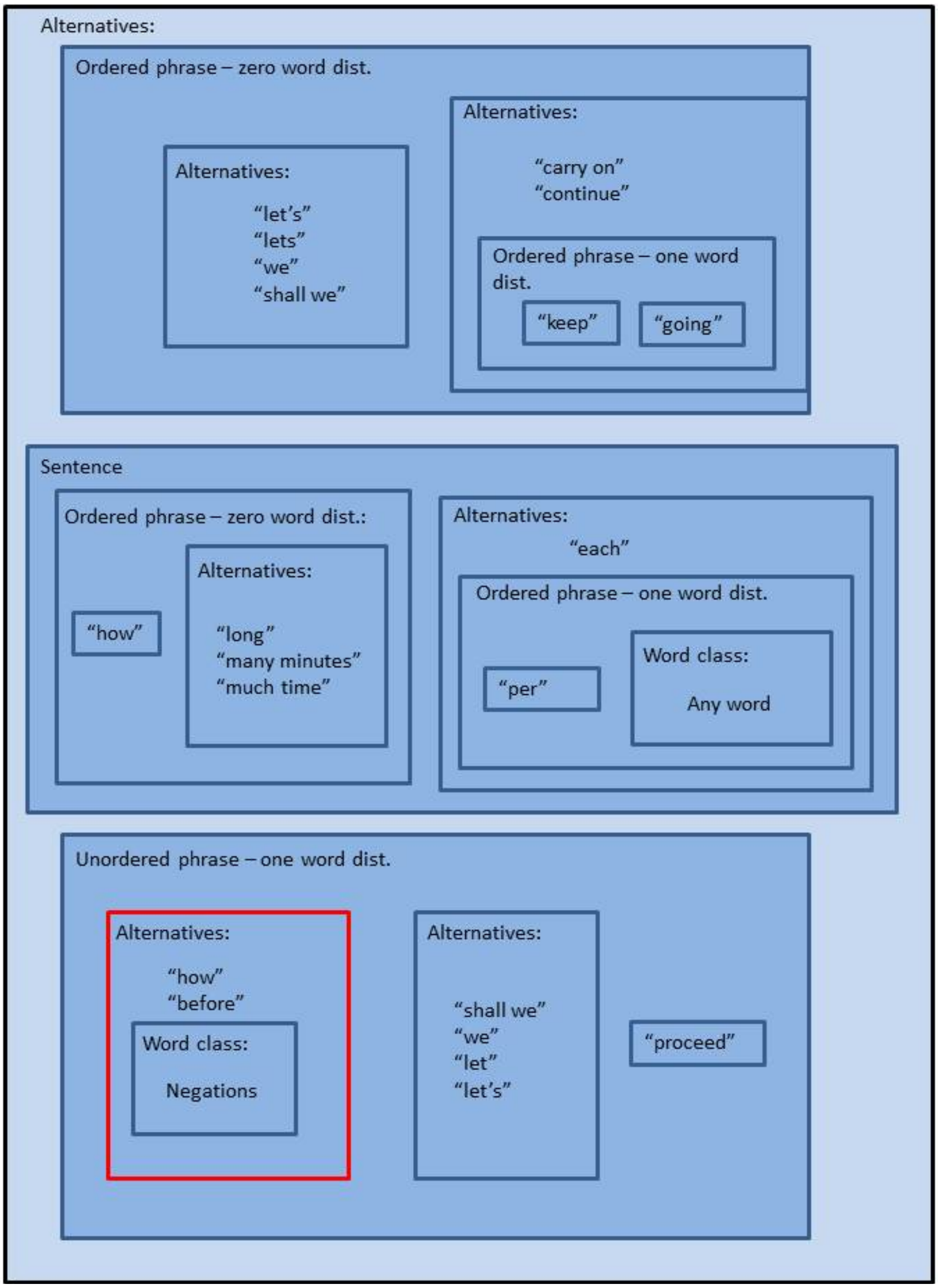


**Figure 3-22 Pacing query part 2**

The first phrase in part 2 of the pacing query, illustrated in Figure 3-22, aimed to pick up further variations of phrases a therapist might use within a session to move the patient onto the next item or part of the appointment. The third phrase in this part of the query constituted yet another expression of this kind of language from the therapist but it was found to be necessary to include a set of negated alternatives. The purpose of this negated item was to exclude from results two types of phrases that would otherwise be picked up by this phrase item. The first of these exclusions was phrases in which a therapist was asking how the patient wished to 'proceed' with either session, treatment or to complete certain tasks prior to the therapy session. For example, 'How should we proceed with this treatment?' or 'Could you complete the worksheet before we proceed.' These phrases generally referred to how the patient would like to do something, for example whether they did want to continue with treatment or with the therapy session that day. These do not fit with the aim of the query, which was to pick up evidence of a therapist keeping a therapy session to time, so were excluded from the results. The second involved the use of negations around the idea of proceeding with treatment. It is likely that in this second case the therapist would be referring to a patient not proceeding with treatment (e,g, 'we shouldn't/won't proceed'), an idea that would not fit within the 'pacing' item. Finally, the second (middle) phrase aimed to count references to timing, these were often expressed in the context of agenda setting when the therapist and patient might have decided together how long to spend on each item, topic or point within the agenda.

#### 3.3.5.6 Interpersonal effectiveness

The interpersonal effectiveness item on the CTS-R scale is made up of three elements: empathy, genuineness and warmth. It refers primarily to the therapist's manner and ability to put the patient at ease and thus to develop an appropriate connection (Blackburn et al., 2001).

Query development for this item was more complex as it was more difficult to classify the language and phrases that would be relevant than for the previous items. The process involved determining common sympathetic phrases that would be appropriate within this context such as 'that must be difficult' or words of encouragement such as 'well done', as well as phrases that might allow the patient to feel understood. Query development was further guided by manual reading of transcripts from the IPCRESS clinical trial (D. Kessler et al., 2009) in order to find other phrases and terms that might provide evidence of interpersonal effectiveness in a therapist. The query is illustrated in Figure 3-23.

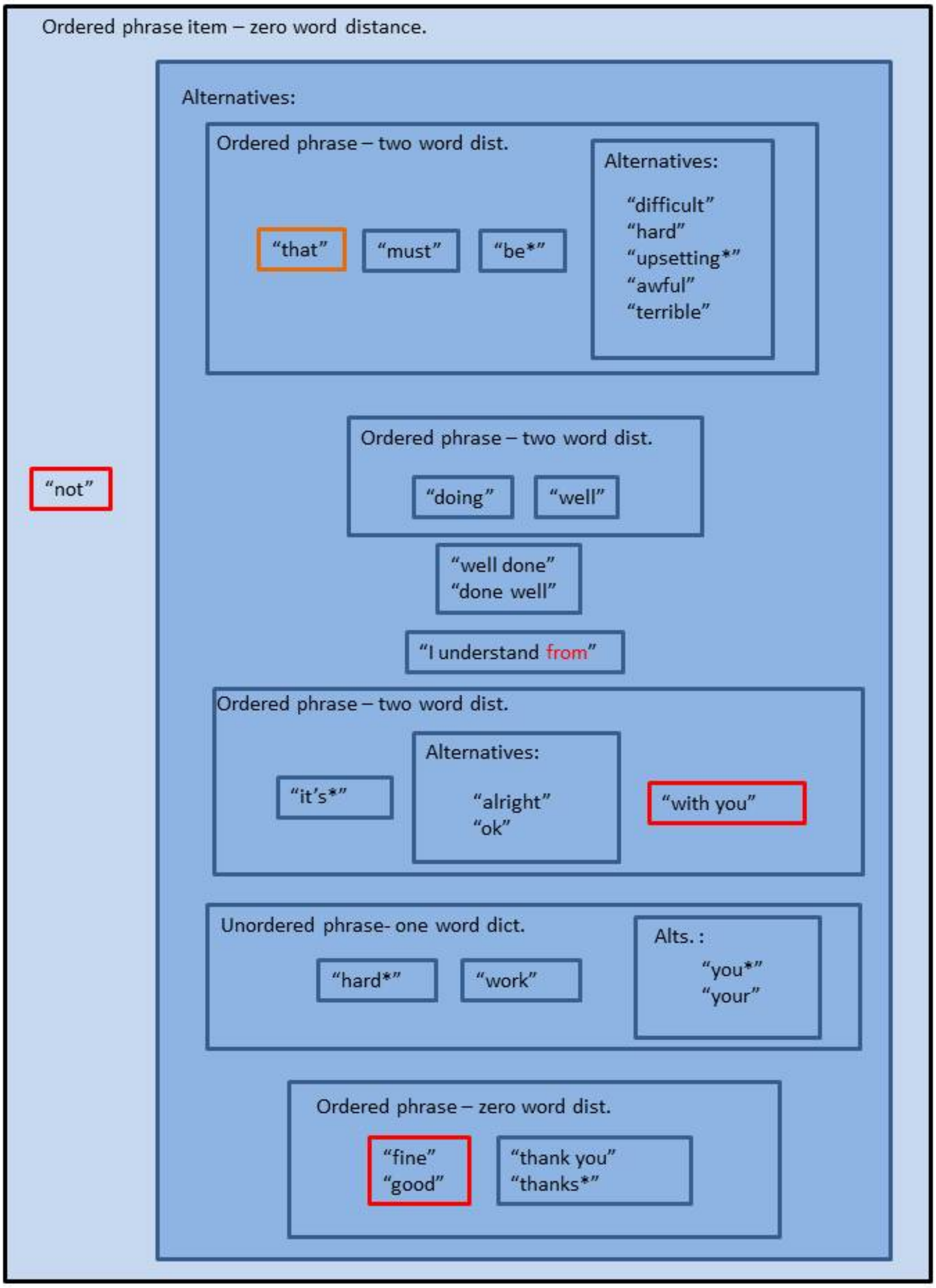


**Figure 3-23 Interpersonal effectiveness query**

It was a large phrase item made up of the negated term 'not' and a set of alternatives. The negated 'not' was a part of this query in order to exclude instances where the therapist may have been expressing the opposite sentiment to that that was intended to be picked up here. Within the set of alternatives, there were seven phrases. The first and fourth phrases aimed to pick up expressions of sympathy from the therapist. This included acknowledging a patient's feeling in a situation with phrases such as 'I understand' or 'that must have been difficult'. In the case of the phrase 'I understand', it was followed by the negated word 'from' so as to exclude phrases such as 'I understand from your assessment' where the focus would be on learning information rather than an expression of sympathy.

The second and third phrases in the larger set of alternatives were phrases of encouragement ('well done') and the fifth phrase aimed to pick up phrases of reassurance 'it's ok'. The sixth phrase aimed to count references towards a patient's 'hard work' and the final, seventh phrase looked to pick up phrases thanking the patient for their input as this was considered a form of encouragement on the part of the therapist. This final phrase included the negated terms 'fine' and 'good' so as to exclude from this count expressions such as 'I'm fine, thank you' that might occur routinely at the beginning of a session.

### 3.4 Linguistic data extraction

Queries for the individual linguistic variables described in the previous section were run on the development set. Results within I2E can be displayed in a variety of ways including by frequency or by document. Displaying results by document provided a frequency score for individual documents. These results were then exported as a Microsoft Excel document. Results for patient and therapist language were exported for each query then transferred to the main results dataset that contained appointment and demographic information as well as the outcome scores listed above. Frequency scores were transformed into proportional scores using word counts. The full dataset could then be imported into STATA in order to perform statistical analyses.

See Table 3-7 for the full list of linguistic variables used within the analysis. All variables listed were counted separately for patients and therapists and a score for each was calculated.

**Table 3-7 Summary table of linguistic variables extracted**

| Label | Origin/source | Description |
|---|---|---|
| **Typing rate** | Calculated from word count and appointment length | Words typed per appointment minute. |
| **LIWC Negative** | LIWC | Proportion of language used that fits into LIWC Negative language category |
| **LIWC Positive** | LIWC | Proportion of language used that fits into LIWC Positive language category |
| **LIWC Social** | LIWC | Proportion of language used that fits into LIWC Social language category |
| **LIWC Certainty** | LIWC | Proportion of language used that fits into LIWC certainty language category |
| **LIWC Insight** | LIWC | Proportion of language used that fits into LIWC Insight language category |
| **LIWC Negations** | LIWC | Proportion of negations used as defined by the LIWC category Negations. |
| **LIWC 'I'** | LIWC | Proportion of First person singular pronouns used as defined by the LIWC category First person singular pronouns. |
| **LIWC 'We'** | LIWC | Proportion of First person plural pronouns used as defined by the LIWC category First person plural pronouns. |
| **I2E Negative** | LIWC-based, developed with I2E | Proportion of negative language used as measured by the LIWC-based query developed in I2E |
| **I2E Positive** | LIWC-based, developed with I2E | Proportion of positive language used as measured by the LIWC-based query developed in I2E |

| | | |
|---|---|---|
| **PANAS-X Negative** | PANAS-X, expanded and developed with I2E | Proportion of negative language used as measured by the query developed in I2E based on the expanded version of the PANAS-X negative language category. |
| **PANAS-X Positive** | PANAS-X, expanded and developed with I2E | Proportion of positive language used as measured by the query developed in I2E based on the expanded version of the PANAS-X positive language category. |
| **PANAS-X Hostility** | PANAS-X, expanded with I2E | Proportion of hostility language used as measured by the expanded version of the PANAS-X hostility language subcategory (Patient only) |
| **PANAS-X Fear** | PANAS-X, expanded with I2E | Proportion of fear language used as measured by the expanded version of the PANAS-X fear language subcategory (Patient only) |
| **PANAS-X Sadness** | PANAS-X, expanded with I2E | Proportion of sadness language used as measured by the expanded version of the PANAS-X sadness language subcategory (Patient only) |
| **PANAS-X Guilt** | PANAS-X, expanded with I2E | Proportion of guilt language used as measured by the expanded version of the PANAS-X guilt language subcategory (Patient only) |
| **PANAS-X Joviality** | PANAS-X, expanded with I2E | Proportion of joviality (joy) language used as measured by the expanded version of the PANAS-X joviality language subcategory (Patient only) |
| **PANAS-X Self-assurance** | PANAS-X, expanded with I2E | Proportion of self-assurance language used as measured by the expanded version of the PANAS-X negative language subcategory (Patient only) |
| **PANAS-X Attentiveness** | PANAS-X, expanded with I2E | Proportion of attentiveness language used as measured by the expanded version of the PANAS-X attentiveness language subcategory (Patient only) |
| **Agenda setting** | CTS-R, query developed with I2E | Proportion of language making reference to or providing evidence for the setting of an agenda by the therapist, based on the CTS-R item Agenda Setting (Therapist only) |

| | | |
|---|---|---|
| **Homework** | CTS-R, query developed with I2E | Proportion of language making reference to or providing evidence for the setting or discussion of homework, based on the CTS-R item Homework (Therapist only) |
| **Pacing** | CTS-R, query developed with I2E | Proportion of language making reference to or providing evidence that a therapist is actively working to pace a session effectively, based on the CTS-R item Pacing (Therapist only) |
| **Interpersonal effectiveness** | CTS-R, query developed with I2E | Proportion of language that provides evidence for a therapist having good interpersonal skills, based on the CTS-R item Interpersonal Effectiveness (Therapist only) |

## 3.5 Statistical Analysis

### 3.5.1 Overview

Statistical analyses were carried out using Stata 12.0. There were three sections to the statistical analysis carried out in this project. The first, making up the bulk of the analysis, involved the use of mixed effects modeling in order to explore and consider the predictive potential of language used within treatment sessions for associated mental health outcome scores (Aim 2.1). Mixed-effects models are a form of regression modeling within which both fixed effects and random effects predictors can be measured. This allows the model to cope with clustered data, which is the case in this data set as it includes repeated measurements from the same patient. The second applied linear regression and considered the predictive value of language use early in treatment for final outcome scores measured at the end of the course of therapy (Aim 2.2). In the third and final section, Cox's regression model was applied in order to look at time to drop-out in the data set and determine whether there is an association between time to drop-out and any of the variables considered (Aim 2.3). After exploration of associations between mental health outcomes and language features, predictive models of outcome scores either during or at the end of treatment were developed. The aim was to develop a model that was able to predict the outcome of interest based on the values of a set of predictor variables. The modelling process involved fitting an appropriate regression model depending on the outcome of interest to the data and then evaluating the model's performance on data that were not used to fit the model.

Prior to providing details of each set of analyses and how these were carried out, it is helpful to describe an approach that was followed for all three sets of analyses. In each branch of analysis, the linguistic variables developed were explored in stages as opposed to all at the same time. A model termed as 'baseline model' in this thesis was developed from demographic details and baseline outcome scores. Models building on this baseline model were then fitted by including the LIWC variables, LIWC-based sentiment, PANAS-X variables and CTS-R variables in turn. This process allowed the exploration of associations between linguistic features and outcome measures and to select candidate variables for the development of predictive models. The variables that were significantly associated with outcome in these individual models were then combined to develop a predictive model for each outcome measure.

A number of factors led to this approach. The exploratory nature of the project means that a high number of variables, both linguistic and not, were put forward as potential predictors throughout the project. The limited number of patients in the dataset meant that all predictors could not be tested together as the associated power would be limited. It was also expected that a number of linguistic variables would be correlated as they consisted of different approaches to measuring the same basic concept, PANAS-X, LIWC and I2E query-based measures of sentiment are the main examples that were likely to be correlated. Multicollinearity within a regression model can lead to the coefficients for individual predictors being unstable so it is preferable to avoid highly correlated predictors being included in the same model. Finally, the linguistic variables were selected and queries developed sequentially meaning that models were developed for each set of language features in turn. Thus, the selection of subsequent variables was in part informed by the results of previous models. For the reasons stated above, it was determined that a logical, sequential approach to developing the statistical models should be followed.

### 3.5.2 Sample size

Throughout the modelling within this project, I aimed to follow the rule of thumb of 10 events or subjects per variable as a guide to the maximum number of predictors that could be included within one model (Harrell, 2001). For example, there were 233 patients who completed treatment in the development data set. This would suggest that a maximum of 23 candidate predictor variables be included when fitting the model. Model development was carried out on a dataset that had been previously extracted (from the service records) and anonymised, meaning that keeping to the suggested ratio of ten cases to a predictor relied on limiting the number of predictors. For the validation set, I looked to include a minimum of 100 events in logistic regression models (Justice, Covinsky, & Berlin, 1999) and maintain the ratio of a minimum of ten cases to each predictor for continuous outcomes (Collins, Ogundimu, & Altman, 2016). Based on this, the validation set was extracted to include data from a minimum of 150 patients who completed treatment. The data set extracted was also set to include data from patients who attended treatment at the same time as those who completed treatment but dropped out of treatment or were referred elsewhere. This provided sufficient data for survival analyses of time to drop-out.

### 3.5.3 Demographic variables for baseline models

All demographic information within the dataset was provided by Ieso Digital Health and the categories and information available are those collected routinely by the service. Gender, age group, step and provisional diagnosis were all included as potential predictors in baseline models along with baseline measures of PHQ-9 and GAD-7 scores. The baseline outcome scores were entered as continuous variables but the four demographic variables were entered as categorical measures. Gender was entered with two categories, and the information was considered missing where this was not specified. Five categories were included for age group: 18-29 years, 30-39 years, 40-49 years, 50-59 years, and over 60 years. In the total data set there were four groups within the Step variable: Assessment, Step 2, Step 3 and Step 3+. However, 'assessment' indicates that no step was allocated, often because a patient did not complete the assessment process with Ieso Digital Health. This means that no transcripts and consequently linguistic data were available for these individuals. In the data that was used for modelling, all patients had been allocated a 'step', meaning that there were only three categories (Steps 2, 3 and 3+) included in the statistical analysis. Provisional diagnoses were provided as a range of diagnostic labels, which were allocated by the GP at the time of referral or adjusted by the therapist after assessment. Given the large number of dummy variables that would required to include all labels as categories in the modelling process, the provisional diagnosis was reduced to three broad categories: 'Anxiety-based diagnoses', 'Depression-based diagnoses', and 'Mixed or Other diagnoses'. These were the three categories included throughout modelling. 'Time' in this data was included in analysis as the number of appointments to date (including the current appointment). Gaps between treatment sessions vary greatly both between and within individuals and therapists so 'number of sessions' was selected as a more appropriate measure of 'time in therapy'. Where this was significant, a squared measure of time in therapy was also included in order to account for any potential non-linear effects of number of sessions on outcome.

### 3.5.4 Mixed effects models

Mixed effects regression models make up the bulk of the analysis within this research project. These are a form of regression modelling that allows the inclusion of random effects in addition to fixed effects. Random effects are included when data points are not expected to be independent and are clustered for any given reason. For example if data is collected from a number of different medical practices or schools, the individuals providing the data are likely to be more similar, thus correlated, within a school or medical practice than across different school or practices. Including clustering information as random effects allows total variance to be split and a relevant portion attributed to the source of clustering in order to avoid misattribution of variance to the tested predictors, or fixed effects, or the overestimation of error within the model.

In the case of the data in this project, there were two potential sources of clustering. The first was the repeated measures design in that each patient attended multiple therapy sessions, each of which had its own set of associated data. The second potential source of clustering was due to therapist identity. There were 661 patients included in the development set and 65 therapists. It is possible that patients were more similar within than between therapists due a variety of possible reasons including how much training and experience a therapist has, their specialty, skill or therapeutic style. For these reasons, both therapist and patient identity were initially included as random effects and these were maintained or removed from the analysis depending on the magnitude of the estimated intra class correlations. The intra class correlation is a measure of how closely data points within a group or cluster are related. It was calculated as the ratio of variance accounted for by the random effect (e.g. between patient variance) over the total variance in the data (between patient variance plus error variance).

Assumptions of normality and homogeneity of variance required by the models were evaluated graphically using residual plots.

#### 3.5.4.1 Outcome scores

Four sets of mixed-effects models were fitted with two versions of the two outcome scores as dependent variables. Each therapy session has a PHQ-9 score and a GAD-7 score associated with it that the patient is requested to complete up to two days before the treatment session. These count as the first version of the outcome score that will be referred to as ‘outcome before session’. A model was developed for each of these scores taken before the session. This model measured the potential association between language use in a treatment session and the outcome score before the session in order to consider whether language features were reflective of mental health outcome, putting linguistic features forwards as possible markers of mental health status and progress of mental health treatment. A model was also developed for each of the scores recorded up to two days before the next session. This is the second version of the outcome score and will be referred to as ‘outcome before next session’. This model focused on the association between language use in a treatment session and the outcome before the next therapy session, considering whether language features were potential short-term predictors of outcome.

#### 3.5.4.2 Predictor variables: measures of linguistic features

In both of the models described above, language features were considered in relation with mental health outcomes measured either before or after the treatment session. The linguistic features defined earlier in this chapter were extracted from individual sessions to form a single score per session.

This same process was followed to extract the measures for each linguistic feature. This consistency then enables models that bring together linguistic features from different sets as candidate predictors.

#### 3.5.4.3 Model development

Models were fitted and predictors selected within the development dataset and the final developed model was later tested on the validation set. A baseline model was first developed that considers time (measured in number of appointments to date), gender, age, provisional diagnosis, step, baseline GAD-7 and PHQ-9 scores, and typing rate (measured as the number of words typed divide by the length of the appointment and expressed in words per minute). Gender, age, provisional diagnosis and step were all included as categorical variables. Building on the baseline model and including significant predictor variables, a model was then developed and cross-validated for each set of linguistic variables. These sets were: 1) the eight LIWC categories described previously, 2) the LIWC-based measures of negative and positive language and Negations, Social language, Insight, Certainty, First person pronouns singular and plural from the LIWC dictionary, 3) the nine PANAS-X-based measures, and 4) the four CTS-R items for which queries were developed. The final, predictive, model included the combination of significant predictors from each set of linguistic variables and was both cross-validated and externally validated using an unseen data set.

Models were developed using a backwards stepwise approach using a significance threshold of 15% since a 15% significance levels has been shown to perform better in variable selection compared to, for example, using the conventional 5% level (Ambler, Seaman, & Omar, 2012). Each set of linguistic variables was entered into the model as a set of potential predictors and predictors that did not reach the 15% level of significance were removed from the model one by one until all included predictors were significant at this level.

#### 3.5.4.4 Cross-validation

Models were internally validated within the development data set using five-fold cross-validation. The aim with cross-validation is to determine how well a model might perform on an independent data set and consequently, practice. For the five folds, 1/5$^{th}$ of the data is reserved as a test set, acting as an ‘unseen’ set while the model is fitted on the remaining 4/5$^{ths}$ of the data. Predictions of values of the outcome in the test set are then made and compared with actual observed values. This is then repeated until each section of the data is used as a test set. Repeating this process five times involved splitting the data into five random groups. Data from one patient across multiple sessions was kept in the same group. The R-squared was estimated for each fold as measures of model fit. R-squared is defined here as one minus the ratio of the variance of the residual values (observed minus predicted values) over the total variance. It gives an estimate of the proportion of variance in the outcome that is explained by the predictors in the model. The R-squared values presented will be estimated considering fixed effects only as opposed to also including random effects as this provides a better estimate of how the model would perform on an independent dataset. Additionally, a calibration slope was estimated. This is the coefficient or slope in a regression analysis of the observed values of the outcome on the predicted values. The closer this score is to 1, the better the agreement between the observed and predicted values is estimated to be.

#### 3.5.4.5 External validation

The combined model made up of the linguistic features from each set of linguistic variables that were significantly associated with outcome was developed and then tested on an external dataset. This was done by estimating the parameters of the model on the development set and predicting outcome scores from these and the linguistic measures in the new data set. The R-squared and calibration slope were then estimated to determine model fit. Fixed-effects residuals were also estimated and their distribution checked graphically to verify that these were distributed normally.

### 3.5.5 Linear regression

Linear regression models were developed in order to analyse the associations between language use at the beginning of therapy and outcome scores at a planned end of treatment. The aim with this set of analyses was to determine whether language used early in treatment can provide an indication of therapy success at a later date (Aim 2.3). Models were developed following the same approach as described within the mixed effects models. These models were developed using only data from patients who had completed their course of treatment and been discharged upon agreement with their therapist.

#### 3.5.5.1 Outcome variables

Two final outcome scores were considered; these were the PHQ-9 and GAD-7 scores at the last treatment session, respectively referred to as ‘End of treatment PHQ-9’ and ‘End of treatment GAD-7’.

#### 3.5.5.2 Predictor variables: Linguistic variables early in treatment

The aim with the linear regression modelling of the end of treatment outcome score was to consider language early in treatment. There was therefore a need to determine what would qualify as ‘early’ in treatment. Within IAPT, attendance at one assessment session and one treatment session qualifies an individual to have ‘engaged’ in therapy and attendance at two therapy sessions qualifies an individual as having ‘completed’ therapy regardless of whether or not they attend the full course of treatment offered (eight sessions on average). Conversely, Ieso Digital Health defines treatment completion as a patient having attended treatment sessions until discharge by mutual agreement with their therapist to end treatment. The data set included in the linear regression analysis includes only those who completed treatment based on the Ieso Digital Health definition and patients would therefore have attended an average of eight sessions before leaving treatment. This definition of treatment completed was selected for this analysis for two reasons. It seems to be a far more common and acceptable length of treatment for a course of CBT and it allowed a time gap between predictor measurement and end of treatment outcome that would potentially make the model useful in practice. If the IAPT definition were followed, there would be cases in the data set in which the time gap between predictor measurement and end of treatment outcome would be only the time between one session and the next.

Given the suggestion that patients could gain enough from two treatment sessions to be considered to have ‘completed treatment’, it seems that these are considered meaningful and it was decided that language used in the first two attended treatment sessions would be used for these regression analyses. It would have been possible to use measures of language used only in the first treatment session as predictors for this analysis. However, mean values across two sessions were selected as these may provide a more typical or representative measure of a patient's language use early in treatment than using measures from a single session.

For each language variable developed, the mean score from the first two attended treatment sessions was calculated to provide an 'early in treatment' score (this does not include the assessment session). These were used as predictor variables in this set of regression modelling.

#### 3.5.5.3 Model development

The linear regression models for this section of the analysis were developed similarly to the mixed effect models. Backwards stepwise regression was used to select predictor variables and 15% was chosen as the significance threshold for inclusion in the model. As with the mixed effects models, a baseline model was initially developed and then the four sets of linguistic models were developed separately and in turn. Significant predictors from each of these models were then included as potential predictors in a final, combined model of final therapy outcome.

#### 3.5.5.4 Model validation

As with the mixed effects models, the linear regression models were both cross-validated and externally validated following the same process (see 3.5.3.3 and 3.5.3.4).

### 3.5.6 Clinical outcomes

#### 3.5.6.1 Logistic regression

Within IAPT, treatment success is based on a binary measure of recovery. A patient is deemed to have recovered from treatment if their end of treatment PHQ-9 score is below 10 and their end of treatment GAD-7 score is below 8.

In order to provide more clinically meaningful models, logistic regression models were fitted for each of these two outcomes. The process for this was almost identical to that followed for the linear regression analyses described in 3.5.5 with the only differences being the binary outcome scores and the estimates of model fit. In the case of the logistic regression models, the c-statistic was estimated as a measure of model performance. These models were also externally validated through estimation of the c-statistic and of a calibration slope. This slope was the coefficient in a logistic regression model fitted with the linear prediction of outcome calculated from the parameters of the developed model as the only covariate and recovery as the binary outcome.

#### 3.5.6.2 Cox proportional hazards model of time to drop-out

The final set of analyses carried out within this project were survival analyses examining risk of drop-out from therapy. A Cox model considers the time at risk of a particular event occurring, in this case the time to a patient dropping out of treatment, and the event occurrence in relation to covariates entered into a model. Based on this information the model developed estimates the hazard (here, risk of dropping out). In this analysis, the hazard ratios associated with covariates will be reported. For example, a hazard ratio of 1 indicates no effect of a covariate on the risk of drop-out and a hazard ratio of 1.3 indicates a 30% increase in the risk of drop out for every unit change in the covariate. The proportional hazards assumption was testing using Schoenfeld residuals.

A Cox model was used here as it allows the inclusion of time-varying covariates in estimating the risk of drop-out. Individual patients were in treatment, and therefore at risk of dropping out, for varying amounts of time. This is a characteristic that a Cox model is designed to handle. Additionally, potential associations between language used in a treatment session and an individual's risk of dropping out of treatment were to be explored, with levels of language feature use changing over time.

The models developed were not used as predictive models but as exploratory models to investigate potential explanatory variables for drop-out from treatment. Internal and external validations were therefore not carried out. Models were developed using backwards stepwise selection as with previous models. Sets of linguistic variables were explored separately before combining significant variables to form a final model within each data set.

#### 3.5.6.2.1 Drop-out as outcome

An individual was deemed to have dropped out of treatment if they did not complete treatment according to the definition of completion used by Ieso Digital Health, described above.

**Table 3-8 Summary table of analyses to be performed**

| Research question | Outcome | Details of candidate predictor variable | Case and session numbers | | | | Model type |
|---|---|---|---|---|---|---|---|
| | | | Development set | | Validation set | | |
| | | | All eligible cases | Completed cases | All eligible cases | Completed cases | |
| Are language features used in therapy sessions markers of mental health state? | PHQ-9 score before session | Baseline (Baseline PHQ-9, Gender, Age, Step, Diagnosis, Number of sessions) | 379 cases - 1906 appts | 206 cases - 1353 appts | N/A | N/A | Mixed effects regression |
| | GAD-7 score before session | Baseline (Baseline GAD-7, Gender, Age, Step, Diagnosis, Number of sessions) | 375 cases- 1883 appts | 201 cases - 1322 appts | N/A | N/A | Mixed effects regression |
| | PHQ-9 score before session | Measure of linguistic features at individual sessions: LIWC; LIWC-based; PANAS-X based; CTS-R based | 374 cases - 1758 appts | 206 cases - 1266 appts | 293 cases - 1138 appts | 173 cases - 900 appts | Mixed effects regression |
| | GAD-7 score before session | | 370 cases - 1741 appts | 201 cases - 1240 appts | 293 cases - 1130 appts | 172 cases - 896 appts | Mixed effects regression |
| Are language features used therapy sessions predictors of short-term | PHQ-9 score before next session | Baseline (Baseline PHQ-9, Gender, Age, Step, Diagnosis, Number of sessions) | 376 cases - 1832 appts | 204 cases - 1286 appts | N/A | N/A | Mixed effects regression |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| mental health outcomes? | GAD-7 score before next session | Baseline (Baseline GAD-7, Gender, Age, Step, Diagnosis, Number of sessions) | 372 cases - 1825 appts | 199 cases - 1272 appts | N/A | N/A | Mixed effects regression |
| | PHQ-9 score before next session | Measure of linguistic features at individual sessions: LIWC; LIWC-based; PANAS-X based; CTS-R based | 372 cases - 1685 appts | 203 cases - 1196 appts | 204 cases - 908 appts | 172 cases - 769 appts | Mixed effects regression |
| | GAD-7 score before next session | | 369 cases - 1683 appts | 200 cases - 1198 appts | 204 cases - 908 appts | 172 cases - 769 appts | Mixed effects regression |
| Are language features used early in treatment predictors of end of treatment mental health outcomes? | End of treatment PHQ-9 score | Baseline (Baseline PHQ-9, Gender, Age, Step, Diagnosis, Total number of sessions) | N/A | 207 cases | N/A | 159 cases | Linear regression |
| | End of treatment GAD-7 score | Baseline (Baseline GAD-7, Gender, Age, Step, Diagnosis, Total number of sessions) | N/A | 207 cases | N/A | 159 cases | Linear regression |
| | End of treatment PHQ-9 score | Mean measures from first two treatment sessions: LIWC; LIWC-based; PANAS-X based; and CTS-R based features | N/A | 207 cases | N/A | 159 cases | Linear regression |
| | End of treatment GAD-7 score | | N/A | 207 cases | N/A | 159 cases | Linear regression |
| Are language features used early in treatment predictors of PHQ-9 based recovery? | PHQ-9 based recovery | Baseline (Baseline PHQ-9, Gender, Age, Step, Diagnosis, Total number of sessions) | N/A | 207 cases | N/A | 159 cases | Logistic regression |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Early in treatment measures of: LIWC; LIWC-based; PANAS-X based; CTS-R based features | N/A | 207 cases | N/A | 159 cases | Logistic regression |
| Are language features used early in treatment predictors of GAD-7 based recovery? | GAD-7 based recovery | Baseline (Baseline GAD-7, Gender, Age, Step, Diagnosis, Total number of sessions) | N/A | 203 cases | N/A | 159 cases | Logistic regression |
| | | Mean measures from first two treatment sessions: LIWC; LIWC-based; PANAS-X based; CTS-R based | N/A | 203 cases | N/A | 159 cases | Logistic regression |
| Are language features used in a therapy session associated with drop-out from treatment? | Time to drop out | Baseline (PHQ-9 and GAD-7 scores, Gender, Age, Step, Diagnosis, Number of sessions attended, typing rate) | 473 cases | N/A | 348 cases | N/A | Cox regression |
| | | Measure of linguistic features at individual sessions: LIWC; LIWC-based; PANAS-X based; CTS-R based | 473 cases | N/A | 348 cases | N/A | Cox regression |

NB. For candidate sets of linguistic features a separate model was first developed to explore associations and then a combined model of candidate predictors was developed. All models including linguistics features within one research question were developed on data from the same number of appointments.

# Chapter 4. Results from Linguistic Inquiry and Word Count measures of language

This chapter presents the results of models developed for each outcome score considered with the selected Linguistic Inquiry and Word Count (LIWC) dictionary categories as potential predictors in the models. In each case, the baseline predictors that had a statistically significant association with outcome were accounted for. Models were fitted on data from all cases included in the complete development data set as well as on a subset of these that included only data from patients who had completed their course of treatment at the time of data collection. The aim in doing this was to consider any population differences between models fitted on the whole data set and a self-selecting data set of individuals who completed therapy. For each outcome considered, the model fitted on the full data set is presented first with any differences in the model when fitted on the subset of completed patients highlighted subsequently. In this and the following chapters, four mixed effects models will be presented covering two versions of two outcome scores: the PHQ-9 and GAD-7 scores recorded before a session or before the next session.

## 4.1 Note on baseline models and variables

Prior to the development of models fitted with linguistic features as candidate predictor variables, a baseline model was developed for each outcome. These included demographic and baseline outcome scores as predictors and any variables that were found to have a statistically significant association with outcome were included in the models presented in this chapter so as to account for these. The full results of the baseline models can be found in the appendix. Below is a list of the variables tested in the baseline models, some of which will be included in the models presented in this chapter.

The demographic variables included within these baseline models were the following:

**Number of sessions:** This is used as a measure of time and is the number of appointments a patient has made up to that point.

**Number of sessions (squared)**

**Baseline PHQ-9:** The PHQ-9 score reported before the first session, which is the assessment session. Mean = 12.60 (SD = 6.46)

**Baseline GAD-7:** The GAD-7 score reported before the first session, which is the assessment session. Mean = 12.00 (SD = 5.29)

**Gender:** Coded as 1 for male, 0 for female, missing if not provided.

**Step group:** This is a categorical variable with three categories that provides an indication of the severity of an individual's mental health condition. These are Step 2, 3 or 3+. In regression results these appear as dummy variables for Step 3 and Step 3+ with Step 2 as the reference category.

**Diagnostic group:** This is the broad diagnostic group to which a patient has been allocated. This was either Depression, Anxiety or Mixed/Other. In regression results, these appear as dummy variables for Anxiety and Mixed diagnostic groups with the Depression diagnostic group as the reference category.

**Age group:** This is a categorical variable with five categories: 18-29 years, 30-39 years, 40-49 years, 50-59 years or 60+ years. These were included in analysis with four dummy variables, with the 18-29 year group as the reference category.

### 4.1.1 Random effects

Both therapist and patient were explored as potential random effects in the mixed effects analyses in this chapter and throughout the project. The value of including both of these as random effects was evaluated by estimating the intra-class correlation of the data points. In an empty model the intra-class correlation for repeated measurements within an individual was 0.71 and the intra-class correlation for therapist was 0.02. The individual was still included as a random effect as the intra-class correlation was approximately 0.35 or 35% after inclusion of demographic information and the first set of linguistic features (LIWC). In this model the intra-class correlation for therapist was below 0.01 and the estimated regression coefficients did not change after excluding therapist identity. It was therefore decided that it was not necessary to include therapist identity as a random effect in the models developed. Though the majority of the drop in the intra-class correlation for therapist was accounted for by baseline and demographic measures, this process was also carried out with the different sets of linguistic features with near-identical results but will not be reported on in subsequent chapters.

For each outcome, the results presented are the coefficients, p-values and 95% confidence intervals associated with the fixed effects predictors included in the model.

## 4.2 Description of candidate predictor variables

Sixteen candidate predictor variables were considered in this section of analysis. These correspond to the patient and therapist measures for each of the following eight LIWC categories: Negative language, positive language, first personal singular pronouns, first person plural pronouns, social language, negations, insight language and certainty language. Table 4-1 presents the descriptive statistics for each of these candidate variables within the development dataset. Note that scores for linguistic features refer to the percentage of language used by each person in a session that counts as the linguistic feature measured. These are calculated separately for patient and therapist as the number of hits for each language feature over the total number of words typed in the session. As can be seen there is quite a large amount of variability in these numbers both within and between linguistic variables with mean percentage of first person pronoun use at 0.42% and mean patient positive language use at 4.74%, for example.

**Table 4-1 Summary statistics for LIWC linguistic features**

| Linguistic feature | Mean percentage score | SD | Min | Max |
|---|---|---|---|---|
| **Patient negative language (LIWC)** | 2.52 | 1.15 | 0 | 8.33 |
| **Patient positive language (LIWC)** | 4.74 | 1.65 | 0 | 16.19 |
| **Patient Social (LIWC)** | 1.08 | 0.86 | 0 | 7.14 |
| **Patient use of negations (LIWC)** | 2.45 | 1.06 | 0 | 13.89 |
| **Patient Insight (LIWC)** | 3.54 | 1.24 | 0 | 8.46 |
| **Patient Certainty (LIWC)** | 1.37 | 0.68 | 0 | 5.17 |
| **Patient first person singular pronouns (LIWC)** | 3.35 | 1.30 | 0 | 11.11 |
| **Patient first person plural pronouns (LIWC)** | 0.42 | 0.45 | 0 | 4.76 |
| **Therapist negative language (LIWC)** | 2.15 | 1.07 | 0 | 7.13 |
| **Therapist positive language (LIWC)** | 5.87 | 1.655 | 1.38 | 14.66 |
| **Therapist Social (LIWC)** | 0.54 | 0.51 | 0 | 4.97 |
| **Therapist use of negations (LIWC)** | 0.94 | 0.56 | 0 | 4.08 |
| **Therapist Insight (LIWC)** | 3.81 | 1.16 | 0 | 8.69 |
| **Therapist Certainty (LIWC)** | 0.21 | 0.14 | 0 | 2.65 |
| **Therapist first person singular pronouns (LIWC)** | 2.14 | 0.95 | 0 | 7.52 |
| **Therapist first person plural pronouns (LIWC)** | 1.38 | 0.73 | 0 | 5.26 |

## 4.3 Model results

### 4.3.1 Outcome 1 – PHQ-9 score before session.

This model was fitted with the PHQ-9 score taken before a session as the outcome score. A total of 1758 observations were used in the model, corresponding to data from 374 patients. Table 4-2 presents the results for the fixed effects predictors included in this model.

**Table 4-2 Results from model predicting PHQ-9 score before session from LIWC linguistic features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.69 | [ 0.63 ; 0.75 ] | <0.001 |
| ***Number of sessions*** | -0.36 | [ -0.42 ; -0.29 ] | <0.001 |
| ***Step group 2 – reference group*** | | | |
| ***Step group 3*** | 0.71 | [ -0.23 ; 1.65 ] | <0.001 |
| ***Step group 3+*** | 3.06 | [ 1.86 ; 4.25 ] | |
| ***Patient Negative language (LIWC)*** | 0.29 | [ 0.11 ; 0.47 ] | <0.001 |
| ***Patient Positive language (LIWC)*** | -0.16 | [ -0.29 ; -0.04 ] | 0.012 |
| ***Patient Negations (LIWC)*** | 0.23 | [ 0.03 ; 0.44 ] | 0.025 |
| ***Patient Social language (LIWC)*** | 0.38 | [ 0.14 ; 0.62 ] | 0.002 |
| ***Patient First person singular pronouns*** | 0.15 | [ -0.04 ; 0.35 ] | 0.122 |
| ***Therapist positive language (LIWC)*** | -0.10 | [ -0.23; 0.03 ] | 0.131 |
| ***Therapist certainty language (LIWC)*** | -1.03 | [ -2.38 ; 0.32 ] | 0.134 |
| ***Therapist negations (LIWC)*** | 0.31 | [ -0.02 ; 0.65 ] | 0.071 |
| ***Constant*** | 1.27 | [ -0.29 ; 2.82 ] | 0.111 |

Baseline PHQ-9 score, number of appointments and step group remained significantly associated with outcome in this model when the linguistic predictors were included. Note that significance levels (p-values) presented for categorical variables (Step and Diagnostic group) were estimated through

Wald tests of the joint significance of the dummy variables for each category. Of the sixteen candidate linguistic predictors put forward in this model, eight were retained in the model at the 0.15 threshold used in this analysis. Five of these predictors were suggested to be positively associated with PHQ-9 score, these were: patient negative language, patient use of first person singular pronouns, patient use of negations, patient use of social language and therapist use of negations. The results therefore suggest that greater use of language fitting within these categories was associated with a higher PHQ-9 score (worse depression outcome) recorded before the session. The three remaining linguistic predictors were negatively associated with PHQ-9 score before a session; patient positive language, therapist positive language and therapist certainty. In the case of these predictors, the results suggest that using a higher proportion of language fitting within these categories is associated with lower PHQ-9 scores and therefore improved depression outcome.

The coefficients associated with these linguistic predictors can be interpreted as follows. In the case of patient negative language use, the associated coefficient was 0.29 (95% CI = [ 0.11 ; 0.47], $p < 0.001$) meaning that for every percentage point increase in negative language use, the associated PHQ-9 score was expected to be, on average, 0.29 points higher. This means that a patient whose language was made up of 4% negative language words during their session, was expected to have scored approximately 0.29 of a point higher on the PHQ-9 score measured before that session than a patient who used only 3% of negative words during their therapy session. Each of the linguistic variables can be interpreted in this way. Negative coefficients suggest a lower PHQ-9 score recorded before a session where more language within a given category was used. For example, in the case of therapist positive language, a session in which 2% of the therapist’s language fits within the positive language LIWC category is associated with a PHQ-9 score recorded before the session that is, on average, 0.1 of a point lower than a session in which only 1% of the therapist’s language can be considered positive.

When the equivalent model was fitted on data from only patients who completed their course of therapy, some differences between the models (in Table 4-2 and Table 4-3) emerged. The model was fitted on data from 1266 appointments involving 206 patients. In the model considering only data from those who completed treatment, the associations between three of the linguistic variables and outcome were not statistically significant at the 15% level when they had been in the same model fitted on a full data set. These three linguistic features were patient first person singular pronouns, therapist positive language and therapist certainty. Additionally, therapist insight language was included in the model when it had not been in the previously presented model. The associated coefficient of 0.16 (95% CI = [-0.02 ; 0.35], $p = 0.089$) suggests that higher levels of therapist insight were associated with a higher PHQ-9 score recorded before the session. However, the p-value attached to this is reaching towards 0.1, suggesting weaker evidence supporting this association than for patient negative language and patient social language, both significantly positively associated with outcome with attached p-values below 0.01. It is also interesting to note that the evidence supporting the positive association between patient negation use and outcome score is much weaker in this model than the previous model.

Details of the results in the dataset containing only data from patients who completed treatment can be found in Table 4-3.

**Table 4-3 Results from model predicting PHQ-9 score before a session - completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.62 | [ 0.54 ; 0.70 ] | < 0.001 |
| ***Number of sessions*** | -0.41 | [ -0.48 ; -0.33 ] | < 0.001 |
| ***Step group 2 – reference group*** | | | |
| ***Step group 3*** | 1.32 | [ 0.10 ; 2.54 ] | < 0.001 |
| ***Step group 3+*** | 3.61 | [ 2.07 ; 5.14 ] | |
| ***Patient Negative language (LIWC)*** | 0.40 | [ 0.18 ; 0.61 ] | < 0.001 |
| ***Patient Positive language (LIWC)*** | -0.15 | [ -0.30 ; 0.01 ] | 0.060 |
| ***Patient Negations (LIWC)*** | 0.18 | [ -0.07 ; 0.44 ] | 0.149 |
| ***Patient Social language (LIWC)*** | 0.44 | [ 0.13 ; 0.75 ] | 0.005 |
| ***Therapist insight language (LIWC)*** | 0.16 | [ -0.02 ; 0.35 ] | 0.089 |
| ***Therapist negations (LIWC)*** | 0.36 | [ -0.04 ; 0.76 ] | 0.077 |
| ***Constant*** | 0.22 | [ -1.59 ; 2.04 ] | 0.808 |

### 4.3.2 Outcome 2 – GAD-7 score before session.

This model was fitted with the GAD-7 score taken before a session as the outcome score. A total of 1741 observations were used in the model, corresponding to data from 370 patients. Table 4-4 presents the fixed effects for this model.

**Table 4-4 Results from model predicting GAD-7 before session**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.61 | [ 0.53 ; 0.68] | <0.001 |
| ***Number of sessions*** | -0.67 | [ -0.87 ; -0.46 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.002 ; 0.03] | 0.024 |
| ***Diagnostic group1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.59 | [ -0.27 ;1.47] | 0.081 |
| ***Diagnostic group 3 (Mixed)*** | 1.11 | [ 0.14 ;2.09] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.92 | [ -0.05 ; 1.89] | <0.001 |
| ***Step group 3+*** | 3.17 | [ 1.94 ; 4.39 ] | |
| ***Patient Negative language (LIWC)*** | 0.31 | [ 0.15 ;0 .49 ] | <0.001 |
| ***Patient Negations (LIWC)*** | 0.20 | [ 0.003 ;0.39 ] | 0.046 |
| ***Patient Social language (LIWC)*** | 0.37 | [ 0.14 ; 0.60 ] | 0.002 |
| ***Patient first person plural pronouns (LIWC)*** | -0.35 | [ -0.75 ; 0.05 ] | 0.086 |
| ***Therapist positive language (LIWC)*** | -0.16 | [ -0.23 ; 0.01 ] | 0.083 |
| ***Therapist negations (LIWC)*** | 0.29 | [ -0.03 ; 0.61 ] | 0.080 |
| ***Therapist certainty (LIWC)*** | -1.36 | [ -2.66 ; -0.07 ] | 0.039 |
| ***Constant*** | 0.22 | [ -1.59 ; 2.04 ] | 0.808 |

As with the previously presented model, the predictors that were retained in the baseline model were also retained here. The coefficients associated with baseline GAD-7 scores, number of appointments, squared number of appointments, diagnostic group and step group were similar to those in the baseline model.

Among the sixteen candidate linguistic predictors, seven were included in the model presented here. When compared with the model predicting PHQ-9 outcome before the session, there are two notable differences. The first is that the association between patient positive language and GAD-7 score is not significantly associated with outcome in this model. Similarly, the association between patient first personal singular pronoun use and outcome score was not retained in this model. However, the association between patient first person plural pronoun use and outcome was.

Patient first person plural pronoun use was suggested to be negatively associated with GAD-7 score measured before the session. The coefficient of -0.35 (95% CI =[ -0.75 ; 0.05 ], $p$ = 0.086) associated with patient use of first person plural pronouns (‘we’, ‘our’, etc.) suggests that for every percentage of first person plural pronouns a patient used during a given treatment session, the GAD-7 score recorded before the session was expected to be 0.35 points lower, on average. However, the evidence behind this association, as suggested by the p-value, was weaker than that supporting the association between patient negative language or patient social language and outcome. Though there were some small differences in the coefficient sizes associated with these and the remaining predictors, these were broadly similar to those presented in the equivalent model predicting PHQ-9 score before the session and consistently with the same direction of association.

When a model was developed with the same outcome score and candidate predictor variables using only data from patients who had completed treatment, the results were quite different. This model was fitted on data from 1240 appointments involving 201 patients. It is important to note that in this, smaller, dataset, gender was statistically significantly associated with outcome in the baseline model with female patients estimated to have a GAD-7 score higher on average than male patients. It was therefore included in this model as a predictor along with the other baseline predictors retained in the model. The results for this model can be found in Table 4-5. In this model only four of the linguistic predictors were statistically significant at the 15% level. These were patient negative language, patient social language, therapist insight language and therapist use of negations. All of these were positively associated with the GAD-7 score recorded before the session. The coefficients associated with patient negative language, patient social language and therapist use of negations varied only slightly from those presented in the model fitted on data from all patients. Though therapist insight language had not been retained in the model fitted on the full data set, it was here and in the equivalent model predicting PHQ-9 score before the session, with almost identical coefficients.

**Table 4-5 Results from model predicting GAD-7 before session – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.56 | [ 0.46 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.74 | [ -0.98 ; -0.50 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.0003 ; 0.03 ] | 0.045 |
| ***Diagnostic group 1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.34 | [ 0.18 ; 2.51 ] | 0.046 |
| ***Diagnostic group 3 (Mixed)*** | 1.38 | [ 0.07 ; 2.70 ] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 0.98 | [ -0.30 ; 2.27 ] | 0.002 |
| ***Step group 3+*** | 2.88 | [ 1.24 ; 4.52 ] | |
| ***Gender*** | 1.20 | [ 0.06 ; 2.34 ] | 0.040 |
| ***Patient Negative language (LIWC)*** | 0.41 | [ 0.21 ; 0.61 ] | <0.001 |
| ***Patient Social language (LIWC)*** | 0.32 | [ 0.03 ; 0.62 ] | 0.031 |
| ***Therapist insight (LIWC)*** | 0.17 | [ -0.01 ; 0.34 ] | 0.065 |
| ***Therapist negations (LIWC)*** | 0.33 | [ -0.05 ; 0.70 ] | 0.090 |
| ***Constant*** | -0.15 | [ -2.14 ; 1.83 ] | 0.882 |

### 4.3.3 Outcome 3 – PHQ-9 score before next session.

This model was fitted with the PHQ-9 score recorded before the next session as outcome or dependent variable. The model used data from 1685 appointments, corresponding to 372 patients. The number of appointments was lower than in the previous model due to the nature of the outcome variable. As it is the outcome score associated with the next session, the last session a patient attends was missing this outcome. The results of this model can be found in Table 4-6.

**Table 4-6 Results from model predicting PHQ-9 score before the next session**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.65 | [ -0.85 ; -0.045 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.005 ; 0.04 ] | 0.012 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.89 | [ -0.07 ; 1.82 ] | <0.001 |
| ***Step group 3+*** | 3.41 | [ 2.22 ; 4.60 ] | |
| ***Patient Negative language (LIWC)*** | 0.17 | [ 0.003 ; 0.35 ] | 0.047 |
| ***Patient Negations (LIWC)*** | 0.16 | [ -0.05 ; 0.36 ] | 0.129 |
| ***Patient Social language (LIWC)*** | 0.20 | [ -0.03 ; 0.43 ] | 0.090 |
| ***Patient first person sign. Pronouns (LIWC)*** | -0.16 | [ -0.35 ; 0.03 ] | 0.105 |
| ***Therapist negations (LIWC)*** | 0.50 | [ 0.15 ; 0.86 ] | 0.006 |
| ***Constant*** | 1.35 | [ 0.01 ; 02.69 ] | 0.048 |

As with previous models, the predictors that were significantly associated with outcome in the baseline models remained so when tested with the linguistic variables included in the model. In addition to these baseline measures, five linguistic variables were retained in the model at the 0.15 threshold. These were patient negative language, patient negation use, patient use of first person singular pronouns, patient use of social language and therapist use of negations. Only patient use of first person singular pronouns was suggested to be negatively associated with the outcome score in this model ($b$ = -0.16, 95% CI = [ -0.35 ; 0.03 ], $p$ =0.105). This stands in contrast to the coefficient associated with the same predictor variable in the model looking at PHQ-9 score before session where the associated coefficient was 0.15 (95% CI = [ -0.04 ; 0.35 ], $p$ = 0.122) and therefore positively associated with outcome. This changing relationship may indicate an unstable predictor or a difference in the nature of the association between patient first person singular pronoun use and PHQ-9 score before and after a therapy session. The p-value around 0.1 also suggests that the evidence supporting these relationships is weak, further putting into question the validity of this association.

The remaining four variables in this model were positively associated with outcome suggesting that a greater presence of these language features in the therapy transcripts was associated with higher PHQ-9 scores measured before the next session. These were patient negative language, patient negation use, patient use of social language and therapist use of negations. These four predictors were included in the model presented in 4.3.1, considering PHQ-9 score before a given session, but coefficients and p-values were generally slightly weaker in this model with the exception of therapist negation use, for which the coefficient went from 0.31 (95% CI = [ -0.02 ; 0.65 ], *p*= 0.071) previously to 0.50 (95% CI = [ 0.15 ; 0.86 ], *p* = 0.006 in this model. As well as suggesting a weaker association between three of these variables and outcome, the evidence supporting the reality of the effect was lower in this model.

When the equivalent model was developed using data from only patients who completed their course of treatment, three of the linguistic predictors mentioned above were retained. The model was fitted on data from 1196 appointments involving 203 patients. Results can be found in Table 4-7. Patient use of first person singular pronouns was negatively associated with outcome score, as was the case in the previous model but the coefficient was stronger, and the p-value lower; -0.28 (95% CI = [ -0.51 ; -0.03 ], *p* = 0.022) here compared to -0.16 (95% CI = [ -0.35 ; 0.03 ], *p* = 0.105 ) in the model covering all patient cases. Patient negative language and patient social language were not significantly associated with outcome in this version of the model. However, the association between patient certainty and outcome, that was not included in any previous versions of models looking at PHQ-9 score before a given or the next session, was retained in the model.

The associated p-value of 0.128 suggests caution in interpretation as it suggests weak evidence supporting this association.

**Table 4-7 Results from model predicting PHQ-9 score before next session – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| *Baseline PHQ-9* | 0.61 | [ 0.53 ; 0.69 ] | <0.001 |
| *Number of sessions* | -0.75 | [ -0.98 ; -0.51 ] | <0.001 |
| *Number of sessions (squared)* | 0.02 | [ 0.004 ; 0.04 ] | 0.012 |
| *Step group 2* | | | |
| *Step group 3* | 1.43 | [ 0.18 ; 2.68 ] | <0.001 |
| *Step group 3+* | 3.84 | [ 2.28 ; 5.40 ] | |
| *Patient Negations (LIWC)* | 0.18 | [ -0.06 ; 0.44 ] | 0.141 |
| *Patient Certainty language (LIWC)* | -0.27 | [ -0.61 ; 0.08 ] | 0.128 |
| *Patient first person sing. Pronouns (LIWC)* | -0.28 | [ -0.51 ; -0.03 ] | 0.022 |
| *Therapist negations (LIWC)* | 0.60 | [ 0.17 ; 1.02 ] | 0.006 |
| *Constant* | 3.03 | [ 1.42 ; 4.64 ] | <0.001 |

### 4.3.4 Outcome 4 – GAD-7 score before next session.

This model was fitted with the GAD-7 score recorded before the next session as the outcome. The model used data from 1683 appointments, corresponding to 369 patients. The results for this model can be found in Table 4-8.

**Table 4-8 Results from model predicting GAD-7 score before next session from LIWC features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.58 | [ 0.51 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.80 | [ -0.96 ;-0.62 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.03 | [ 0.01 ; 0.04 ] | <0.001 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.68 | [ -0.008 ; 2.37 ] | 0.067 |
| ***Diagnostic group 3 (Mixed)*** | 1.34 | [ 0.17 ; 2.10 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.22 | [ 0.26 ; 2.19 ] | <0.001 |
| ***Step group 3+*** | 3.63 | [ 2.42 ; 4.84 ] | |
| ***Patient negations (LIWC)*** | 0.15 | [ -0.04 ; 0.34 ] | 0.116 |
| ***Therapist Negative language (LIWC)*** | 0.29 | [ 0.12 ; 0.46 ] | 0.001 |
| ***Therapist Positive language (LIWC)*** | -0.17 | [ -0.30 ; 0.04 ] | 0.012 |
| ***Therapist negations (LIWC)*** | 0.25 | [ -0.09 ; 0.59 ] | 0.144 |
| ***Constant*** | 2.33 | [ 0.82 ; 3.85 ] | 0.003 |

The predictors that were significantly associated with outcome in the baseline version of this model remained so when the model was fitted here with the linguistic predictors included. Four linguistic predictors were statistically significant at the 15% level in this model, these were: patient and therapist use of negations and therapist use of negative and positive language. Three of these were also included in the model looking at GAD-7 score before a session and the association was in the same direction. Therapist negative language was the additional predictor in the model, which was positively and significantly associated with outcome score ($b$ = 0.29, 95% CI = [ 0.12 ; 0.46 ], $p$ = 0.001). This suggests that for every percentage of negative language used by a therapist in an appointment, the GAD-7 score attached to the following appointment was expected to be an average of 0.29 of a point higher. Both patient and therapist use of negations were also suggested to be positively associated with outcome, which was also the case in the model predicting PHQ-9 score before the next session. However, the significance values associated with both of these suggest there is only weak evidence supporting the associations. Only therapist positive language was negatively and significantly associated with outcome (*b* = - 0.17, 95% CI = [ -0.30 ; 0.04 ] , *p* = 0.012 ) in this model suggesting that greater use of positive language by the therapist in a session was associated with a lower GAD-7 score measured before the next session.

When the equivalent model was fitted on only data from patients who completed their course of treatment, there was only one primary difference with the model presented above. Therapist use of negations was not included in this model (included in appendix - 0). Patient use of negations and therapist positive and negative language were all retained in the model with coefficients of association with outcome in the same directions as in the previous model but with some small increases in the magnitude of the estimated coefficient. This model was fitted on data from 1198 appointments involving 200 patients.

The summary table of linguistic predictors included previously (Table 4-1) suggests that though there is quite a bit of variability between linguistic features, the numbers associated remain quite low and in single digits in the majority of cases. This means that, though some associations between predictors and outcome scores put forward were statistically significant, sometimes highly so, both the coefficients and proportions of language used were generally quite low, suggesting that only a small proportion of the outcome score was being explained by each linguistic feature. The results from the cross-validation presented in Table 4-10 can be seen to support this idea as the increase in the mean R-squared is low when compared to baseline model cross-validation summary statistics that have been included here for easy comparison.

### 4.3.5 Cross-validation

**Table 4-9 Summary results from five-fold cross-validation of baseline models.**

| *Outcome* | *All or completed cases* | *Mean cross-validated $R^2$* | *Range of $R^2$* | *Calibration slope* | *Intercept* |
|---|---|---|---|---|---|
| ***Outcome 1 – PHQ-9 before session*** | All cases | 0.50 | [ 0.37 – 0.64] | 0.97 | 0.38 |
| | Completed only | 0.41 | [ 0.30 – 0.61] | 0.95 | 0.55 |
| ***Outcome 2 – GAD-7 before session*** | All cases | 0.36 | [0.28 – 0.52] | 0.93 | 0.68 |
| | Completed only | 0.32 | [ 0.26 – 0.39] | 0.93 | 0.80 |
| ***Outcome 3 – PHQ-9 before next session*** | All cases | 0.47 | [0.34 – 0.62] | 0.96 | 0.50 |
| | Completed only | 0.38 | [ 0.24 – 0. 61] | 0.93 | 0.82 |
| ***Outcome 4 – GAD-7 score before next session*** | All cases | 0.34 | [ 0.23 – 0.49] | 0.92 | 0.91 |
| | Completed only | 0.30 | [ 0.20 – 0.37] | 0.91 | 1.03 |

In this table providing results from the cross-validation of baseline models The strongest model in terms of the reported mean R-squared was that predicting the PHQ-9 score reported before a therapy session. A mean R-squared of 0.50, suggesting that it explains 50% of the variation in PHQ-9 scores, puts forward a strong model, even with only baseline variables included. The model predicting PHQ-9 score reported before the next session reported a slightly weaker mean R-squared. In the case of models predicting GAD-7 outcome scores, the associated R-squared measures were on average 0.10 weaker than those associated with the equivalent PHQ-9 models. Similarly, for each outcome put forward, the model developed using the full data set seemed slightly stronger than that developed using only data from patients who completed their course of treatment. This may, however, be associated with the number of data points included.

**Table 4-10 Summary results from five-fold cross-validation**

| Outcome | All or completed cases | Mean cross-validated $R^2$ | Range of $R^2$ | Calibration slope | Intercept |
|---|---|---|---|---|---|
| ***Outcome 1 – PHQ-9 before session*** | All cases | 0.52 | [ 0.41 – 0.64] | 0.98 | 0.16 |
| | Completed only | 0.45 | [ 0.34 – 0.62] | 0.97 | 0.32 |
| ***Outcome 2 – GAD-7 before session*** | All cases | 0.39 | [0.28 – 0.54] | 0.95 | 0.48 |
| | Completed only | 0.34 | [ 0.23 – 41] | 0.94 | 0.62 |
| ***Outcome 3 – PHQ-9 before next session*** | All cases | 0.49 | [0.37 – 0.63] | 0.95 | 0.53 |
| | Completed only | 0.40 | [ 0.30 – 0. 62] | 0.94 | 0.67 |
| ***Outcome 4 – GAD-7 score before next session*** | All cases | 0.36 | [ 0.22 – 0.53] | 0.92 | 0.84 |
| | Completed only | 0.34 | [ 0.19 – 0.46] | 0.92 | 0.86 |

Table 4-10 shows the summary information for the cross-validation carried out for each model described and presented so far. The mean R-squared measured through cross-validation varies greatly depending on the outcome score considered. In both the baseline models and those presented in this chapter, the model looking to predict PHQ-9 score before a session led to the strongest cross-validated R-squared. In the case of this model, the mean R-squared was 0.52, suggesting that this model accounted for 52% of the variation in the data studied. This is slightly stronger than the equivalent model including only baseline predictor variables for which the mean R-squared was 0.50. All calibration slope scores were above 0.90, suggesting acceptable calibration of the model. A calibration slope of 1 would indicate a perfectly calibrated model.

There are two patterns to note in the results of the cross-validation. Looking at the R-squared results, and as mentioned previously, models predicting PHQ-9 score before a session were the strongest, followed by PHQ-9 score before the next session, followed by GAD-7 score before a session and finally, the weaker model was that predicting GAD-7 score before the next session. This is a pattern that stands across both the baseline and LIWC models and in models using data from all patients or only those that have completed treatment. The second pattern to note involves the comparison of the baseline and LIWC cross-validated R-squared values. As described above, in each case the model including LIWC predictor variables is associated with a slightly higher cross-validated R-squared than the model including only baseline variables, suggesting additional variation explained by the LIWC based linguistic features but only a small amount.

### 4.3.6 Outcome 5 – End of treatment PHQ-9 score

This model considered levels of eight LIWC categories within both therapist and patient language in the first two treatment sessions and their potential association with the PHQ-9 score reported at the end of treatment. This model was fitted on data from 207 patient cases. The results are presented below in Table 4-11

**Table 4-11 Results of linear regression predicting final PHQ-9 score from baseline features and linguistic features early in treatment**

| *Final PHQ-9 score* | *b* | *95% CI* | |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.47 | [ 0.36 ; 0.58 ] | <0.001 |
| ***Patient Positive language (LIWC)*** | -0.41 | [ -0.96 ; 0.13 ] | 0.138 |
| ***Patient Negations (LIWC)*** | 0.94 | [ 0.03 ; 1.84 ] | 0.042 |
| ***Patient Social language (LIWC)*** | 0.90 | [ -0.09 ; 1.87 ] | 0.073 |
| ***Therapist positive language (LIWC)*** | -0.46 | [ -0.95; 0.03 ] | 0.068 |
| ***Constant*** | 2.79 | [ -1.12 ; 6.70 ] | 0.161 |

The results suggest that, in addition to the baseline measure of PHQ-9 score, four of the LIWC measures were retained in this model. Three of these were patient language features and one was a feature of therapist language. Patient positive language use early in therapy was negatively associated with final PHQ-9 score with a coefficient of -0.42 (95% CI = [-0.96 ; 0.13 ], $p$ = 0.138). This suggests that a 1% change in mean patient positive language use (LIWC) in the first two treatment sessions was associated with a 0.41 change, on average, in PHQ-9 score reported at the end of treatment. Both patient negation use and patient use of social language were positively associated with outcome score, suggesting that higher levels of these language features early in therapy were associated with a higher PHQ-9 score at the end of treatment. The final predictor in this model was therapist positive language. This variable was negatively associated with outcome with a coefficient of -0.46 (95% CI = [-0.95; 0.03], $p$ = 0.068). This suggests that a 1% higher mean proportion of positive language in therapist language in the first two treatment sessions was associated with an average of a 0.46 point lower PHQ-9 score reported at the end of treatment, and therefore a better depression outcome. Despite the inclusion of these predictors in the model, the associations with outcome were supported by variable and often high significance values, suggesting weak evidence supporting the reality of theses associations.

This model was estimated to account for 34% of the variation in the outcome. This compares with 28.8% of the variation that was estimated to be explained by the baseline model. This model therefore suggests an improvement on the baseline model even if this additional explained variation was small.

### 4.3.7 Outcome 6 – End of treatment GAD-7 score

This model considers levels of eight LIWC categories within both therapist and patient language in the first two treatment sessions and their potential association with the GAD-7 score reported at the end of treatment. This model was fitted on data from 203 patient cases. The results of this analysis are presented below in Table 4-12.

**Table 4-12 Results of linear regression predicting final GAD-7 score from baseline features and linguistic features early in treatment**

| ***Final GAD-7 score*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.40 | [ 0.29 ; 0.51 ] | <0.001 |
| ***Patient Positive language (LIWC)*** | -0.84 | [ -1.36 ; -0.31 ] | 0.002 |
| ***Patient Negations (LIWC)*** | 1.19 | [ 0.39 ; 2.01 ] | 0.004 |
| ***Patient Social language (LIWC)*** | 1.10 | [ 0.24 ; 1.97 ] | 0.013 |
| ***Patient first person singular pronouns (LIWC)*** | 0.44 | [ -0.14 ; 1.02 ] | 0.138 |
| ***Therapist Negations (LIWC)*** | -1.07 | [ -2.30 ; 0.167 ] | 0.090 |
| ***Therapist positive language (LIWC)*** | -0.39 | [ -0.83; 0.04 ] | 0.078 |
| ***Constant*** | 3.06 | [ -0.54 ; 6.66 ] | 0.095 |

The results of this model suggest that six of the tested linguistic features in this model were retained in this model. Therapist language positive language and use of negations and both were suggested to be negatively associated with outcome. This suggests that higher mean levels of these features in therapist language in the two first treatment sessions was associated with a lower GAD-7 score reported prior to the final treatment session. The patient language features that were included in this model were social language, first person singular pronouns, use of negations and positive language. Positive language was negatively and significantly associated with outcome with a coefficient of -0.84 (95% CI = [ -1.36 ; -0.31 ], *p* = 0.002). This suggests that a 1% increase in mean patient positive language use early in therapy was associated with an average of a 0.84 point lower GAD-7 score reported at the end of treatment. Patient first person singular pronoun use, negation use and social language were all positively associated with outcome suggesting that higher levels of these linguistic features used early in treatment were associated with a higher GAD-7 score at the end of treatment, and therefore a worse anxiety outcome. However, the p-values attached to each of these associations suggest that there is stronger evidence supporting the associations between patient social language and patient negations, and outcome than that between patient first person singular pronoun use and outcome.

This model was estimated to explain 32.6% of the variation in the data, compared to 20.4% of the variation explained in the baseline model. This suggests a reasonable increase in the amount of variation explained and therefore a useful addition to the model.

## 4.4 Overview of results

The results put forward in this chapter suggest that overall, a handful of the LIWC variables were suggested to be statistically significant predictors of outcome score whether this was reported immediately before a session, before the next session or at the end of the course of treatment. In each model that was tested on both the full data set and a set of completed cases only, a wider pool of predictors was retained in the model in the larger dataset. This may suggest greater variability and differences in language use but may also be an effect of the larger population. A number of the associations put forward in the models had attached p-values around 0.1 or above. Despite being included in the model, these higher p-values suggest that the evidence support these effects is modest and often weak. However, a subset of these predictors recurred more frequently in the models with very low p-values, suggesting strong evidence of an association between these and outcome. These were negation use, measures of patient positive and negative language and social language use. These were all measures that had been indicated as associated with mental health state in a range of previous work by Pennebaker and others who have worked with the LIWC dictionary. Additionally, the direction of associations with outcome was consistent across the variables in this stronger subset. Despite this, it is not clear what the nature of the suggested associations is. A causal link cannot be directly established, though differences in the models considering outcomes at different time points may provide some insight in this. Though investigating causality was not a primary aim of this project, it is interesting to consider the relationships these results suggest. The presence of an association between patient positive language and PHQ-9 score before a session and the absence of an association of this same language feature with PHQ-9 score before the next session may suggest that using positive language in treatment is more reflective of recent mental health state than predictive of mental health state the following week. Positive language use early in treatment was associated with end of treatment outcome score, perhaps suggesting that a more positive attitude in treatment at the beginning of the course leads to better outcomes. However, it is also possible that a less severe patient will have better end of treatment outcomes and use more positive language early in treatment. It is also likely that many of the associations here are bi-directional. The potential nature of these associations will be further discussed in the last chapter of this thesis.

# Chapter 5. Results from models fitted with I2E measures of affect based on LIWC categories.

This chapter presents the models developed with the four I2E query-based measures of sentiment that were developed based on the categories found in the Linguistic Inquiry and Word Count dictionary. As was the case in the previous chapter, each model was developed with the inclusion of the significant baseline predictors.

## 5.1 Description of candidate predictor variables

Four predictor variables were considered in this section of the analysis. These were therapist and patient measures of positive and negative language as measured by the negative and positive language queries developed in I2E using the methods described in the Chapter 3. Descriptive statistics for each of these variables are included in Table 5-1. The unit of measurement for the linguistic predictors was the proportion of language within a given session transcript that fits within a category as defined by the query (negative or positive language). E.g. percentage of patient language that was negative.

**Table 5-1 Descriptive statistics for LIWC-based I2E linguistic features**

| Linguistic feature | Mean percentage score | St. Dev | Min | Max |
|---|---|---|---|---|
| **Patient negative language (LIWC-based I2E query)** | 1.59 | 0.80 | 0 | 5.92 |
| **Patient positive language (LIWC-based I2E query)** | 2.17 | 0.93 | 0 | 8.33 |
| **Therapist negative language (LIWC-based I2E query)** | 1.29 | 0.77 | 0 | 5.42 |
| **Therapist positive language (LIWC-based I2E query)** | 2.27 | 0.87 | 0 | 10.20 |

## 5.2 Model results

### 5.2.1 Outcome 1 – PHQ-9 score before session.

This model looked at the associations between the PHQ-9 score recorded before a therapy session and the I2E query-based measures developed from the LIWC negative and positive language categories. The model did not include language and outcome score from the first appointment as this was an assessment session and the outcome score attached to this appointment was used as the baseline outcome score. The model was fitted on data from 1758 therapy sessions from 374 individual patients.

**Table 5-2 Results from model predicting PHQ-9 score before session from LIWC-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***p*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.69 | [ 0.63 ; 0.75 ] | <0.001 |
| ***Number of sessions*** | -0.38 | [ -0.44 ; -0.32] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.84 | [ -0.09 ; 1.78 ] | <0.001 |
| ***Step group 3+*** | 3.33 | [ 2.14 ; 4.52 ] | |
| ***Patient negative language (LIWC-based I2E query)*** | 0.30 | [ 0.06 ; 0.59 ] | 0.015 |
| ***Patient positive language (LIWC-based I2E query)*** | -0.48 | [ -0.69 ; -0.26 ] | <0.001 |
| ***Constant*** | 2.70 | [ 1.54 ; 3.86 ] | <0.001 |

The results of this model (Table 5-2) suggest that the baseline predictor variables maintain their position as strong predictors of PHQ-9 score before a session when the I2E query-based measures of sentiment were included as predictors. Of the four candidate linguistic predictor variables tested, only the two patient measures of affective language were statistically associated with outcome in the final model. Patient negative language was significantly positively associated with outcome with a coefficient of 0.30 (95% CI = [0.06 ; 0.59], $p$ = 0.015). This suggests that every percentage of a patient's language that was negative in a given therapy session was associated with an average of a 0.3 of a point higher PHQ-9 score recorded before the session. Patient positive language was negatively associated with outcome with a coefficient of -0.48 (95% CI = [ -0.69 ; -0.26 ], $p < 0.001$), suggesting that for every percentage of a patient's language in a given session that was positive, the associated PHQ-9 score was 0.48 of a point lower, on average.

When the equivalent model was developed on a data set that was limited to the patients who completed their course of therapy, results were very similar. The model was fitted on data from 1266 appointments involving 206 patients. Measures of both positive and negative patient language were statistically significant with very similar coefficients. Given that the model was almost identical, it was not presented here but can be found in the appendix.

### 5.2.2 Outcome 2 – GAD-7 score before session.

This model looked to predict the GAD-7 score recorded before a therapy session from the I2E query-based measures of sentiment of language within the session. As was the case for the previous model, data from the first appointment was not included for the development of this model. Data from 1741 appointments was used in the development of this model, from 370 individual patients.

**Table 5-3 Results from model predicting GAD-7 score before a session from LIWC-based linguistic features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.60 | [ 0.53 ; 0.68 ] | <0.001 |
| ***Number of sessions*** | -0.68 | [ -0.88 ; -0.48 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.02 ; 0.03 ] | 0.022 |
| ***Diagnostic group (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.77 | [ -0.12 ; 1.66 ] | 0.064 |
| ***Diagnostic group 3 (Mixed)*** | 1.14 | [0.15 ; 2.12 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.98 | [ 0.0001 ; 1.96 ] | <0.001 |
| ***Step group 3+*** | 3.34 | [ 2.10 ; 4.58 ] | |
| ***Patient negative language (LIWC-based I2E query)*** | 0.25 | [ 0.009 ; 0.48 ] | 0.042 |
| ***Patient positive language (LIWC-based I2E query)*** | -0.22 | [ -0.43 ; -0.018 ] | 0.033 |
| ***Therapist negative language (LIWC-based I2E query)*** | 0.41 | [ 0.16 ; 0.66 ] | 0.001 |
| ***Constant*** | 2.48 | [ 1.08 ; 3.88 ] | 0.001 |

The results of the model (Table 5-3) suggest that, as with the previous model, the predictors that were significantly associated with outcome in the baseline model also were in this case.

In addition to the baseline predictors, three of the four linguistic measures tested were retained in this model and statistically significantly associated with outcome. As in the previous model, patient negative and positive language were both significantly associated with outcome. The association between patient positive language and GAD-7 score before the session was smaller here with a coefficient of -0.22 (95% CI = [ -0.43 ; -0.018 ], $p$ = 0.033), than that presented in the previous model (-0.48, 95% CI = [ -0.69 ; -0.26 ], $p$ < 0.001). Additionally, therapist negative language was positively associated with outcome with a coefficient of 0.41 (95% CI = [0.16 ; 0.66], $p$ = 0.001), suggesting that higher levels of negative language in the patient and therapist language were associated with higher GAD-7 scores before the session. The individual coefficients in this model can be interpreted in the same way as the other models described. For example, in the case of therapist negative language, the coefficient attached to this predictor was 0.41 (95% CI = [0.16 ; 0.66], *p* = 0.001). This suggests that a session in which the therapist used 2% positive language was associated with a GAD-7 score (taken before the session) that was, on average, 0.41 points higher than the GAD-7 score recorded before a session in which the therapist used only 1% negative language.

When the equivalent model was fitted on the dataset containing only data from patients who completed their course of therapy, one major difference appeared. This model (Table 5-4) was developed on data from 1250 appointments from 202 individual patients. Of the four linguistic predictors tested in this part of the analysis, only patient positive language and therapist negative language were retained in the model. As compared to the model tested on the full data set, the association between patient negative language and outcome was not significant in this model.  The coefficients associated with patient positive language (*b* = -0.23, 95% CI = [-0.47 ; -0.01 ] , *p* = 0.063 ) and therapist negative language (*b* = 0.46, 95% CI = [ 0.17 ; 0.74 ], *p* = 0.002) were very similar in this model as in the previous model but in the case of patient positive language, the higher p-value suggests there is less evidence behind this association in this smaller dataset. The coefficients associated with the baseline predictors were identical.

**Table 5-4 Results from model predicting GAD-7 score before a session from LIWC-based linguistic features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.57 | [ 0.46 ; 0.67 ] | <0.001 |
| ***Number of sessions*** | -0.75 | [ -0.99 ; -0.52 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.001 ; 0.03 ] | 0.043 |
| ***Diagnostic group (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.30 | [ 0.13 ; 2.47 ] | 0.057 |
| ***Diagnostic group 3 (Mixed)*** | 1.35 | [0.02 ; 2.67 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.95 | [ -0.35 ; 2.25 ] | <0.001 |
| ***Step group 3+*** | 3.08 | [ 1.44 ; 4.72 ] | |
| ***Patient positive language (LIWC-based I2E query)*** | -0.23 | [ -0.47 ; -0.01 ] | 0.063 |
| ***Therapist negative language (LIWC-based I2E query)*** | 0.46 | [ 0.17 ; 0.74 ] | 0.002 |
| ***Constant*** | 2.97 | [ 1.24 ; 4.71 ] | 0.001 |

### 5.2.3 Outcome 3 – PHQ-9 score before next session.

This model looked at the associations between the PHQ-9 score recorded before the next session and the I2E measures of sentiment in the language used in the current session. It was developed using data from 1685 appointments from 372 individual patients.

**Table 5-5 Results from model predicting PHQ-9 score before the next session from LIWC-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.64 | [ -0.84 ; -0.45 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.005 ; 0.04 ] | 0.007 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.87 | [ -0.09 ; 1.81 ] | <0.001 |
| ***Step group 3+*** | 3.49 | [ 2.29 ; 4.67 ] | |
| ***Patient positive language (LIWC-based I2E query)*** | -0.22 | [ -0.43 ; 0.01 ] | 0.055 |
| ***Therapist positive language (LIWC-based I2E query)*** | -0.23 | [ -0.45 ; 0.02 ] | 0.057 |
| ***Constant*** | 3.11 | [ 1.83 ; 4.05 ] | <0.001 |

The results (found in Table 5-5) suggest that both patient and therapist measures of positive language statistically significant at the 15% level. Both therapist and patient positive language were suggested to be negatively associated with the outcome score suggesting that higher levels of positive language used by the patient and the therapist were associated with a lower PHQ-9 score before the next session, and thus a better depression outcome.

When the equivalent model was developed on a dataset containing data from patients who completed their course of treatment and were discharged after agreement with their therapist, the model structure and coefficients attached to individual predictors were almost identical. The model was fitted on data from 1196 appointments involving 203 patients. Both therapist and patient positive language were suggested to be negatively associated with outcome (respectively, $b$ = -0.22 (95% CI = [ -0.50 ; 0.05 ] , $p$ = 0.114 ) and $b$ = -0.25 (95% CI = [-0.52 ; 0.02], $p$ = 0.071)) but the higher p-value suggest less confidence in this being a true association. Given how similar these two models were, the table reporting the results of this model will not be included here but can be found in the appendix.

### 5.2.4 Outcome 4 – GAD-7 score before next session.

This model looked to predict the GAD-7 score recorded before the next session from the I2E measures of sentiment in the language used in the current session. It was developed using data from 1683 appointments from 369 individual patients.

**Table 5-6 Results from fixed effect model predicting GAD-7 score before next session from LIWC-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.59 | [ 0.51 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.80 | [ -0.99 ; -0.63 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.03 | [ 0.01 ; 0.04 ] | <0.001 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.71 | [ -0.17 ; 1.59 ] | 0.067 |
| ***Diagnostic group 3 (Mixed)*** | 1.15 | [0.18 ; 2.12 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.21 | [ 0.24 ; 2.18 ] | <0.001 |
| ***Step group 3+*** | 3.71 | [ 2.50 ; 4.93 ] | |
| ***Therapist positive language (LIWC-based I2E query)*** | -0.37 | [ -0.59 ; -0.15 ] | 0.001 |
| ***Therapist negative language (LIWC-based I2E query)*** | 0.26 | [ 0.01 ; 0.50 ] | 0.039 |
| ***Constant*** | 3.09 | [ 1.74 ; 4.43 ] | <0.001 |

The results in Table 5-6 suggest that all predictors that were significantly associated with outcome in the baseline version of the model also were in this version of the model with similar coefficient and significance values. Of the four linguistic predictors tested in this model, only those relating to therapist language were retained and significantly associated with outcome. This is a difference compared to the previous model predicting PHQ-9 score before the next session in which patient positive language had been significantly associated with outcome and therapist negative language had not been. Therapist negative language was positively associated with outcome score with a coefficient of 0.26 (95% CI = [0.01 ; 0.50 ], $p$=0.04) suggesting that for every percent negative language used by the therapist in a session, the GAD-7 score before the next session was expected to be an average of 0.26 of a point higher. Conversely, therapist positive language was negatively associated with outcome. This suggests that a higher level of positive language in a therapy session was associated with a lower GAD-7 score measured before the next session.

The equivalent model was developed on a dataset made up only of data from patients who completed their course of treatment. This model was fitted on data from 1198 appointments involving 200 individuals. There was a difference in the linguistic predictors that were found to be significantly associated with outcome (see Table 5-7) when compared both to the model fitted on the full data set and the model predicting PHQ-9 score before the next session. In all three of these models, therapist positive language was significantly associated with outcome, with a larger coefficient in the models predicting GAD-7 score. Patient negative language was however significantly and positively associated with outcome here suggesting that higher levels of patient negative language were associated with a higher GAD-7 score reported before the next treatment session.

**Table 5-7 Results from model predicting GAD-7 score before next session from LIWC-based features – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.53 | [ 0.543 ; 0.63 ] | <0.001 |
| ***Number of sessions*** | -0.84 | [ -1.06 ; -0.63 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.03 | [ 0.01 ; 0.04 ] | 0.004 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.18 | [ 0.04 ; 2.33 ] | 0.064 |
| ***Diagnostic group 3 (Mixed)*** | 1.37 | [0.08 ; 2.67 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.30 | [ 0.03 ; 2.58 ] | <0.001 |
| ***Step group 3+*** | 3.48 | [ 1.88 ; 5.07 ] | |
| ***Therapist positive language (LIWC-based I2E query)*** | -0.41 | [ -0.66 ; -0.16 ] | 0.002 |
| ***Patient negative language (LIWC-based I2E query)*** | 0.27 | [-0.02 ; 0.55 ] | 0.065 |
| ***Constant*** | 3.30 | [ 1.61 ; 4.99 ] | <0.001 |

### 5.2.5 Cross-validation

Table 5-8 provides a summary of measures estimated from the five-fold cross-validation that was carried out with each of the models described above.

In this table, the same patterns were seen as were found in the cross-validation scores presented in the previous set of results. Models using the full dataset and looking to predict PHQ-9 score taken before the session had the strongest associated mean R-squared from cross-validation. Models predicting PHQ-9 score were consistently stronger than the equivalent models looking to predict GAD-7 scores, models predicting outcome before a session were stronger than models predicting outcome before the following session, and models working with the full set of data were stronger than models working with only data from patients who had completed their course of treatment.

**Table 5-8 Summary results from five-fold cross-validation**

| Outcome | All or completed cases | Mean cross-validated $R^2$ | Range of $R^2$ | Calibration slope | Intercept |
|---|---|---|---|---|---|
| **Outcome 1 – PHQ-9 before session** | All cases | 0.52 | [ 0.41 – 0.65] | 0.98 | 0.15 |
| | Completed only | 0.44 | [ 0.36 – 0.64] | 0.97 | 0.31 |
| **Outcome 2 – GAD-7 before session** | All cases | 0.36 | [0.28 – 0.52] | 0.93 | 0.54 |
| | Completed only | 0.34 | [ 0.25 – 0.46] | 0.95 | 0.56 |
| **Outcome 3 – PHQ-9 before next session** | All cases | 0.49 | [0.37 – 0.62] | 0.96 | 0.46 |
| | Completed only | 0.40 | [ 0.29 – 0. 61] | 0.94 | 0.67 |
| **Outcome 4 – GAD-7 score before next session** | All cases | 0.36 | [ 0.22 – 0.53] | 0.92 | 0.89 |
| | Completed only | 0.33 | [ 0.18 – 0.47] | 0.91 | 0.88 |

### 5.2.6 Outcome 5 – Final PHQ-9 score

This model considers sentiment as measured by LIWC-based sentiment queries developed in I2E during the first two treatment sessions and their association with the PHQ-9 score reported prior to the final treatment session. The variables included in this and the following model refer to mean linguistic feature scores from the first two treatment sessions. This model was fitted on data from 207 patient cases.

**Table 5-9 Results of linear regression predicting final PHQ-9 score from baseline features and LIWC-based linguistic features early in treatment**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.49 | [ 0.39 ; 0.60 ] | <0.001 |
| ***Therapist positive language (LIWC-based I2E query)*** | -1.26 | [-2.22 ; -0.30 ] | 0.011 |
| ***Patient positive language (LIWC-based I2E query)*** | -0.94 | [ -2.02 ; 0.14 ] | 0.087 |
| ***Constant*** | 5.75 | [ 2.76 ; 8.74 ] | <0.001 |

The regression results are presented in Table 5-9. Of the four measures of affect tested in this model, two were retained in the model. These were therapist positive language and patient positive language. Both were suggested to be negatively associated with outcome, suggesting that higher mean positive language early in therapy from both patient and therapist was associated with a lower PHQ-9 score reported prior to the last treatment session. Therapist positive language was significantly associated with outcome with a coefficient of -1.26 (95% CI = [-2.22 ; -0.30 ], $p$ = 0.011). This suggests that a 1% difference in mean therapist positive language in the first two treatment sessions was associated with an average of a 1.26 point difference in the final PHQ-9 score.

This model was estimated to explain 33% of the variation in the outcome scores. This compares to 29% explained by the model including only baseline PHQ-9 score as a predictor variable.

### 5.2.7 Outcome 6 – Final GAD-7 score

This model considered positive and negative language as measured by LIWC-based queries developed in I2E during the first two treatment sessions and their association with the GAD-7 score reported prior to the final treatment session. This model was fitted on data from 203 patient cases.

**Table 5-10 Results of linear regression predicting final GAD-7 score from baseline features and LIWC-based linguistic features early in treatment**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.42 | [ 0.30 ; 0.54 ] | <0.001 |
| ***Therapist positive language (LIWC-based I2E query)*** | -0.96 | [-1.83; -0.09 ] | 0.031 |
| ***Patient positive language (LIWC-based I2E query)*** | -0.90 | [ -1.88 ; 0.07 ] | 0.070 |
| ***Constant*** | 5.10 | [ 2.22 ; 7.99 ] | 0.001 |

The model results (Table 5-10) suggest that the same two linguistic features that retained in the previous model also were here. These were patient and therapist measures of positive language as measured by LIWC-based I2E queries. As in the previous model, both were suggested to be negatively associated with outcome, suggesting that higher levels of therapist and patient language in the first two treatment sessions were associated with a lower GAD-7 score at the end of treatment. Higher levels of positive language early in treatment were therefore suggested to be associated with improved anxiety outcomes at the end of treatment. The attached p-values suggest that there is stronger evidence supporting the presence of an association between therapist positive language and outcome than between patient positive language and outcome.

This model was estimated to explain 24.5% of the variation in the outcome scores. This compares with 20.4% explained by the baseline model. The amount of additional variation explained in this model is almost identical to that in the previous model.

## 5.3 Overview of results

The results presented in this chapter suggest that throughout the models presented, outcome scores were statistically associated with expressions of sentiment as measured by the developed I2E queries. There was however some variation as to whether this was primarily positive or negative language or within the therapist of patient language. Whereas perhaps patient language was more closely associated with outcome score for outcomes 1 and 2, in particular outcome 1 (PHQ-9 score before session), it seems that therapist language played a greater role in models predicting outcomes 3 and 4, particularly outcome 4 (GAD-7 before next session). This was not a clear pattern however, as when only completed cases were included in the model predicting GAD-7 before the next session, patient negative language appeared to be significantly associated with outcome when this had not been the case in the analysis including all patient cases. In the case of final outcome score from language use early in treatment however, it seems clear that the strongest linguistic predictors was therapist positive language, with the suggestion that patient positive language may also be associated with outcome. This suggests that positivity in both parties early in treatment was associated with a stronger likelihood of therapy success and positive language was associated with lower final outcome scores. The direction and mechanisms behind this association are unclear, however. It is possible that positivity from a therapist early in treatment leads to greater engagement of the patient in treatment and therefore improved outcomes, but it is also possible that the nature of a patient's mental health issues influences the positivity and confidence expressed by a therapist about the course of treatment.

Another interesting result from these models was the statistically significant association between patient positive and negative language and PHQ-9 score before a session. This association was also suggested to be present in the model of GAD-7 before a session but with less statistical support. In the models of outcome before the next session however, the evidence supporting this association was weaker, excluding them from the models most of the time. This may suggest that this measure of patient positive and negative language is reflective of mental health state as opposed to predictive of short-term future outcome. The nature of these relationships will be further discussed in the last chapter.

# Chapter 6. Results from models fitted with PANAS-X based linguistic features

This chapter presents the models developed with the predictor variables based on PANAS-X measures of feeling and emotion. These are language features extracted using I2E queries based on the expanded PANAS-X language categories.

## 6.1 Description of the predictor variables

Eleven candidate predictor variables were included in the analysis presented in this chapter. These were therapist and patient measures of both positive and negative language, and seven further measures of patient positive and negative language. These were hostility, guilt, fear and sadness within the negative category and joviality, self-assurance and attentiveness within the positive category. The PANAS-X categories are based on a narrower dictionary than the LIWC categories and, though there is likely to be some overlap, there is a major difference in the size of the two main categories (positive and negative language) as well as in the focus of the subcategories. For example, the LIWC dictionary negative language category includes 740 terms, whereas the equivalent list of expanded PANAS-X terms use here includes only 125 terms. They aim to measure similar concepts, but with a different approach, one using a broader range of terms and the other, a narrower range.

Table 6-1 presents summary measures of the candidate predictor variables considered in this chapter.

**Table 6-1 Summary statistics for Expanded PANAS-X based linguistic features**

| Linguistic feature | Mean percentage score | St. Dev | Min | Max |
|---|---|---|---|---|
| **Patient negative language (Expanded PANAS-X-based I2E query)** | 0.87 | 0.57 | 0 | 6.45 |
| **Patient positive language (Expanded PANAS-X-based I2E query)** | 1.01 | 0.58 | 0 | 4.76 |
| **Patient Joviality (Expanded PANAS-X category)** | 1.21 | 0.71 | 0 | 6.90 |
| **Patient Self-assurance (Expanded PANAS-X category)** | 0.39 | 0.38 | 0 | 5.15 |
| **Patient Attentiveness (Expanded PANAS-X category)** | 0.14 | 0.20 | 0 | 1.83 |
| **Patient Hostility (Expanded PANAS-X category)** | 1.21 | 0.71 | 0 | 6.90 |
| **Patient Guilt (Expanded PANAS-X category)** | 0.03 | 0.10 | 0 | 1.50 |
| **Patient Sadness (Expanded PANAS-X category)** | 0.56 | 0.47 | 0 | 2.95 |
| **Patient Fear (Expanded PANAS-X category)** | 0.44 | 0.49 | 0 | 5.97 |
| **Therapist negative language (Expanded PANAS-X-based I2E query)** | 0.60 | 0.45 | 0 | 3.24 |
| **Therapist positive language (Expanded PANAS-X-based I2E query)** | 1.18 | 0.60 | 0 | 4.36 |

## 6.2 Model results

### 6.2.1 Outcome 1 – PHQ-9 score before session.

This model looked at the associations between the PHQ-9 score attached to a therapy session and the language features extracted from it. The outcome score is therefore the PHQ-9 score reported just before the session. This model was fitted on data from 1758 appointments attended by 374 individual patients. The results can be found in Table 6-2.

**Table 6-2 Results from model predicting PHQ-9 score before session from PANAS-X based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.63 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.35 | [ -0.42 ; -0.29] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.78 | [ -0.15 ; 1.71 ] | <0.001 |
| ***Step group 3+*** | 3.22 | [ 2.04 ; 4.39 ] | |
| ***Patient negative language (Expanded PANAS-X-based I2E query)*** | 0.30 | [ -0.03 ; 0.63 ] | 0.075 |
| ***Patient positive language (Expanded PANAS-X-based I2E query)*** | -0.37 | [ -0.70 ; -0.04 ] | 0.027 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.48 | [ -0.79 ; -0.19 ] | 0.002 |
| ***Therapist positive language (Expanded PANAS-X-based I2E query)*** | -0.61 | [ -0.92 ; -0.30 ] | <0.001 |
| ***Constant*** | 3.53 | [ 2.40 ; 4.67 ] | <0.001 |

The baseline features in this model were maintained as significant predictors of outcome. In addition to the baseline measures, this model suggests that four of the eleven candidate predictor variables were retained in the model. These were patient negative and positive language, patient joviality and therapist positive language. Of these three predictors, only patient negative language was positively associated with outcome (*b* = 0.30, 95% CI = [ -0.03 ; 0.63 ], *p* = 0.075) suggesting that higher levels of negative language were generally associated with a higher PHQ-9 score. However, the higher p-value attached to this predictor suggests there is less evidence supporting its association with outcome than there is for the three remaining predictors. These were negatively and significantly associated with outcome, suggesting that higher levels of patient and therapist positive language and patient joviality were associated with a lower PHQ-9 score measured just before the therapy session and therefore a better depression outcome. Taking the example of patient joviality, the coefficients attached to these predictors can be interpreted as follows. Patient joviality has an associated coefficient of -0.48 (95% CI = [ -0.79 ; -0.19 ], *p* = 0.002) suggesting that for every percentage of patient language that fits within the joviality category, the PHQ-9 score before the session was expected to be an average of 0.49 of a point lower.

When the equivalent model was fitted on data from only patients who completed their course of treatment, there were some differences in the linguistic predictors that were statistically significant predictors of outcome. The model was fitted on data from 1266 appointments from 206 individuals. The associations of therapist positive language and patient joviality with outcome were very similar to those presented in the model fitted on the full data set above. Additionally, patient sadness (Expanded PANAS-X category) was statistically significant at the 15% level. It was suggested to be positively associated with outcome but there was less statistical evidence supporting the presence of this association than the other predictors in model. Results for this model can be found in Table 6-3.

**Table 6-3 Results from model predicting PHQ-9 score before session from PANAS-X based linguistic features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.62 | [ 0.54 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.41 | [ -0.49 ;  -0.34] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.23 | [ 0.002 ; 2.46 ] | <0.001 |
| ***Step group 3+*** | 3.62 | [ 2.07 ; 5.17 ] | |
| ***Patient sadness language (Expanded PANAS-X category)*** | 0.38 | [ -0.10 ; 0.86 ] | 0.122 |
| ***Therapist positive language (Expanded PANAS-X-based I2E query)*** | -0.64 | [ -1.00 ; -0.28 ] | <0.001 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.49 | [ -0.83 ; -0.15 ] | 0.005 |
| ***Constant*** | 3.62 | [ 2.21 ; 5.01 ] | <0.001 |

### 6.2.2 Outcome 2 – GAD-7 score before session.

This model looked at the GAD-7 outcome score recorded before a therapy session based on the expanded PANAS-X features measured during that therapy session. This model was fitted on data from 1741 appointments from 370 individuals.

**Table 6-4 Results from results for model predicting GAD-7 score before session from PANAS-X-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.60 | [ 0.53 ; 0.68 ] | <0.001 |
| ***Number of sessions*** | -0.65 | [ -0.85 ; -0.44 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.02 ; 0.03 ] | 0.022 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.65 | [ -0.23 ; 1.53 ] | 0.087 |
| ***Diagnostic group 3 (Mixed)*** | 1.09 | [0.11 ; 2.067 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.91 | [ -0.07 ; 1.88 ] | <0.001 |
| ***Step group 3+*** | 3.26 | [ 2031 ; 4.49 ] | |
| ***Therapist negative language (Expanded PANAS-X based I2E query)*** | 0.40 | [ -0.003 ; 0.80 ] | 0.052 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.54 | [ -0.83 ; -0.25 ] | <0.001 |
| ***Patient negative language (Expanded PANAS-X based I2E query)*** | 0.39 | [ 0.06 ; 0.72 ] | 0.020 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.47 | [ -0.74 ; -0.19 ] | 0.001 |
| ***Constant*** | 3.54 | [ 2.18 ; 4.90 ] | <0.001 |

The results of this model (Table 6-4) suggest that four of the eleven candidate predictors were statistically significant at 15% in this model. There were two main differences between this model and that predicting PHQ-9 score before the session. These were the absence of a statistically significant association between patient positive language and outcome in this model and the presence of a statistically significant association between therapist negative language and outcome that was not present in the equivalent PHQ-9 model. Therapist negative language was positively associated with outcome suggesting that higher levels of this feature present in the therapy session were associated with a higher GAD-7 score and therefore a worse anxiety outcome. In this model therapist negative language was associated with outcome with a coefficient of 0.40 (95% CI = [ -0.003 ; 0.80 ] , $p$ = 0.052) suggesting that every percentage of negative language used by the therapist during a treatment session was associated with a 0.40 point increase, on average, in GAD-7 score before the session. Patient negative language, patient joviality and therapist positive language were also association with outcome with similar associations as were reported in the previous model.

When the equivalent model was fitted on data from only patients who completed treatment there was one major difference in that patient negative language was not significantly associated with outcome in this model. Therapist positive and negative language were both significantly associated with outcome in the same direction of association and with similar sized coefficients as were presented in the previous model. The association between patient joviality and outcome was also statistically significant in this model with a similar coefficient and the same direction of association as in the previous model. The small changes in coefficient values can be seen in Table 6-5, but aside from these, the model was almost the same as that in Table 6-4. The model was fitted on data from 1250 appointments involving 202 patients.

**Table 6-5 Results from model predicting GAD-7 score before session from PANAS-X-based linguistic features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.56 | [ 0.45 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.69 | [ -0.93 ; -0.45 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ -0.002 ; 0.03 ] | 0.083 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.21 | [-0.53 ; 2.03 ] | 0.070 |
| ***Diagnostic group 3 (Mixed)*** | 1.29 | [1.26 ; 4.49 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.75 | [ -0.53 ; 2.03 ] | <0.001 |
| ***Step group 3+*** | 2.87 | [ 1.26 ; 4.49 ] | |
| ***Therapist negative language (Expanded PANAS-X based I2E query)*** | 0.45 | [ -0.01 ; 0.91 ] | 0.054 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.68 | [ -1.02 ; -0.34 ] | <0.001 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.53 | [ -0.85 ; -0.21 ] | 0.001 |
| ***Constant*** | 3.54 | [ 2.18 ; 4.90 ] | <0.001 |

### 6.2.3 Outcome 3 – PHQ-9 score before next session.

This model looked to predict the PHQ-9 score recorded before the next therapy session from the expanded PANAS-X features extracted from the current therapy session. The model was fitted on data from 1685 appointments from 372 individual patients. Results from this model can be found in Table 6-6.

**Table 6-6 Results from model predicting PHQ-9 score before next session from PANAS-X-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.63 | [ -0.83 ; -0.43 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.005 ; 0.04 ] | 0.010 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.79 | [ -0.16 ; 1.73 ] | <0.001 |
| ***Step group 3+*** | 3.44 | [ 2.25 ; 4.62 ] | |
| ***Therapist negative language (Expanded PANAS-X based I2E query)*** | 0.35 | [ -0.07 ; 0.76 ] | 0.105 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.45 | [ -0.80 ; -0.10 ] | 0.011 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.40 | [ -0.69 ; -0.10 ] | 0.008 |
| ***Constant*** | 3.13 | [ 1.96 ; 4.30 ] | <0.001 |

Both therapist positive and negative language were retained in this model and both were statistically significant at the 15% level. This is a result that contrasts with the first model described in this section where patient, not therapist, negative and positive language were statistically significantly associated with outcome (Table 6-2). Similarly to the previous models presented in this section, patient joviality language was significantly associated with outcome. Therapist positive language and patient joviality were negatively associated with outcome, suggesting that higher levels of therapist positive language ($b$ = -0.45, 95% CI = [ -0.80 ; -0.10 ], $p$ = 0.011) and patient joviality ($b$ = -0.39, 95%CI = [ -0.69 ; -0.10 ], $p$ = 0.008) in a given session were associated with a lower PHQ-9 score before the next session.

The equivalent model was tested on the smaller dataset of 1196 appointments involving 203 patients who completed treatment. The same linguistic predictors were retained in the model. The direction of association was maintained for all three predictors with therapist positive language and patient joviality language negatively associated with outcome and therapist negative language positively associated with outcome. The coefficient associated with patient joviality was stronger, going from -0.40 (95% CI = [ -0.69 ; -0.10 ], *p* =0.008) in the previous model to -0.55(95% CI = [ -0.91 ; -0.19 ], *p* = 0.003) in this model. Therapist positive language was statistically and significantly associated with outcome with a coefficient of - 0.44 (95% CI = [ -0.86 ; -0.03 ], *p* =0.037), suggesting that for every percent of positive language used by a therapist in a given therapy session, the PHQ-9 score taken before the next session was expected to be 0.44 of a point lower, on average. The evidence supporting a positive association between therapist negative language and outcome was weaker than for the other two predictors included in the model.

**Table 6-7 Results from model predicting PHQ-9 score before next session from PANAS-X-based linguistic features – completed cases only.**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.60 | [ 0.52 ; 0.68 ] | <0.001 |
| ***Number of sessions*** | -0.67 | [ -0.91 ; -0.43 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.002 ; 0.04 ] | 0.035 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.33 | [ 0.89 ; 2.58 ] | <0.001 |
| ***Step group 3+*** | 3.70 | [ 2.15 ; 5.25 ] | |
| ***Therapist negative language (Expanded PANAS-X based I2E query)*** | 0.41 | [ -0.10 ; 0.92 ] | 0.116 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.44 | [ -0.86 ; -0.03 ] | 0.037 |
| ***Patient Joviality (Expanded Panas-X category)*** | -0.55 | [ -0.91 ; -0.19 ] | 0.003 |
| ***Constant*** | 3.67 | [ 2.18 ; 5.17 ] | <0.001 |

### 6.2.4 Outcome 4 – GAD-7 score before next session.

This model looked to predict the GAD-7 outcome score before the next session based on language used in the current session. The model was based on data from 1683 appointments from 369 individuals.

**Table 6-8 Results from model predicting GAD-7 score before next session from PANAS-X-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.59 | [ 0.51 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.78 | [ -0.96 ;  -0.60 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.03 | [ 0.01 ;0.04 ] | 0.001 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.65 | [ -0.22 ; 1.53 ] | 0.078 |
| ***Diagnostic group 3 (Mixed)*** | 1.10 | [0.13 ; 2.07 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.10 | [ 0.13 ; 2.07 ] | <0.001 |
| ***Step group 3+*** | 3.59 | [ 2.38 ; 4.80 ] | |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.64 | [ -0.96 ; -0.31 ] | <0.001 |
| ***Therapist negative language (Expanded PANAS-X based I2E query)*** | 0.32 | [ -0.07 ; 0.72 ] | 0.112 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.31 | [ -0.58 ; -0.03 ] | 0.031 |
| ***Constant*** | 3.51 | [ 2.21 ; 4.82 ] | <0.001 |

The results (Table 6-8) suggest that beyond the baseline predictors, three of the candidate linguistic predictor variables tested in this model were retained. As in the previous model, these were therapist positive and negative language and patient joviality language. Patient joviality and therapist positive language were negatively associated with outcome, suggesting that higher levels of these linguistic features in a therapy session were associated with a lower GAD-7 score before the next session, indicative of a better anxiety outcome. As was the case previously, the attached p-values suggest stronger evidence supporting the association between patient joviality and outcome than that between therapist negative language and outcome.

When the equivalent model was tested on a dataset consisting only of data from patients who had completed their course of treatment, two of the predictors included in the model stayed the same, while two others differed. This model was fitted on data from 1198 appointments involving 200 patients. The results can be found in Table 6-9. Patient negative and therapist positive language were significantly associated with outcome in this model. Patient positive language and patient joviality were also included in the model but with evidence supporting the association as indicated by the p-values. Both were suggested to be negatively associated with outcome suggesting that higher levels of these features in a given therapy session were associated with a lower GAD-7 score, and therefore better anxiety outcome measured before the next session.

**Table 6-9 Results from model predicting GAD-7 score before next session from PANAS-X-based features – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.53 | [ 0.43 ; 0.63 ] | <0.001 |
| ***Number of sessions*** | -0.80 | [ -1.02 ; -0.59 ] | <0.001 |
| ***Number of sessions (squared*** | 0.02 | [ 0.006 ; 0.04 ] | 0.008 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.12 | [ -0.02 ; 2.26 ] | 0.088 |
| ***Diagnostic group 3 (Mixed)*** | 1.26 | [ -0.02 ; 2.53 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.13 | [ 0.13 ; 2.40 ] | <0.001 |
| ***Step group 3+*** | 3.21 | [ 1.63 ; 4.79 ] | |
| ***Patient negative language (Expanded PANAS-X based I2E query)*** | 0.43 | [ 0.03 ; 0.83 ] | 0.035 |
| ***Patient positive language (Expanded PANAS-X based I2E query)*** | -0.29 | [ -0.68 ; 0.09 ] | 0.143 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.31 | [ -0.66 ; 0.04 ] | 0.082 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.62 | [ -1.00 ; -0.24 ] | 0.002 |
| ***Constant*** | 3.93 | [ 2.26 ; 5.61 ] | <0.001 |

### 6.2.5 Cross-validation results

Table 6-10 presents the summary statistics associated with the five-fold cross-validation carried out on each of the models presented in this chapter. The pattern and values of these measures are very similar to those found in the previous two chapters with a small amount of additional variation in outcome scores being explained by these models as compared to baseline models. Calibration slope estimates were very similar.

**Table 6-10 Summary results from five-fold cross-validation**

| Outcome | All or completed cases | Mean cross-validated $R^2$ | Range of $R^2$ | Calibration slope | Intercept |
|---|---|---|---|---|---|
| **Outcome 1 – PHQ-9 before session** | All cases | 0.53 | [ 0.41 – 0.67] | 0.98 | 0.15 |
| | Completed only | 0.45 | [ 0.36 – 0.63] | 0.97 | 0.31 |
| **Outcome 2 – GAD-7 before session** | All cases | 0.39 | [0.28 – 0.54] | 0.95 | 0.48 |
| | Completed only | 0.35 | [ 0.28 – 0.46] | 0.96 | 0.46 |
| **Outcome 3 – PHQ-9 before next session** | All cases | 0.49 | [0.37 – 0.62] | 0.96 | 0.49 |
| | Completed only | 0.40 | [ 0.27 – 0. 60] | 0.94 | 0.68 |
| **Outcome 4 – GAD-7 score before next session** | All cases | 0.35 | [ 0.21 – 0.52] | 0.90 | 0.99 |
| | Completed only | 0.33 | [ 0.17 – 0.46] | 0.91 | 0.89 |

### 6.2.6 Outcome 5 – Final PHQ-9 score

This model considered levels of the expanded PANAS-X based linguistic features in language early in therapy and their relationship with the PHQ-9 score reported at the end of the course of treatment. This model was fitted on data from 207 patient cases.

**Table 6-11 Results of linear regression predicting final PHQ-9 score from baseline features and PANAS-X-based linguistic features early in treatment**

| ***Final PHQ-9 score*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.46 | [ 0.28 ; 0.52 ] | <0.001 |
| ***Patient positive language (Expanded PANAS-X-based I2E query)*** | -1.57 | [-3.39; -0.26 ] | 0.092 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -1.71 | [ -3.13 ; -0.29 ] | 0.018 |
| ***Constant*** | 4.92 | [ 2.45 ; 7.38 ] | <0.001 |

The results of the model, presented in Table 6-11, suggest that in addition to the baseline PHQ-9 score, two of the expanded PANAS-X based features were retained in the model. These were the expanded joviality category and the PANAS-X query based measure of positive language, both measured within patient language. Both were negatively associated with outcome suggesting that higher mean levels of these linguistic features early in treatment were associated with a lower PHQ-9 score at the end of the course of treatment. The results can be interpreted similarly as in previous chapters. Joviality, as measured by the expanded PANAS-X category was associated with outcome with a coefficient of -1.7 (95% CI = [ -3.13 ; -0.29 ], $p$ = 0.018). This suggests that a 1% higher mean proportion of Joviality language in the first two treatment sessions was associated with an average of a 1.7 point higher PHQ-9 score at the end of treatment. Of the two associations put forwards, this had the strongest supporting evidence according to the attached p-values.

The variation in outcomes explained by this model was estimated to be 33.4%. This compares with 28.8% in the baseline model. This is a small improvement on the baseline model.

### 6.2.7 Outcome 6 – Final GAD-7 score

This model considers levels of the expanded PANAS-X based linguistic features in language early in therapy and their associations with the GAD-7 score reported at the end of the course of treatment. This model was fitted on data from 203 patient cases.

**Table 6-12 Results of linear regression predicting final GAD-7 score from baseline features and PANAS-X-based linguistic features early in treatment**

| *Final GAD-7 score* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.40 | [ 0.28 ; 0.52 ] | <0.001 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -2.04 | [-3.20; -0.88 ] | 0.001 |
| ***Therapist positive language (Expanded PANAS-X-based I2E query)*** | -1.39 | [ -2.91 ; 0.13 ] | 0.073 |
| ***Constant*** | 5.23 | [ 2.57 ; 7.90 ] | <0.001 |

The results of this model (Table 6-12) suggest that two of the linguistic features developed based on the PANAS-X retained in the model. These were patient joviality as measured by the expanded PANAS-X dictionary and therapist positive language as measured by an I2E query based on the expanded PANAS-X category. Both were negatively associated with outcome score suggesting that higher levels of these linguistic features early in treatment were associated with a lower GAD-7 score reported at the end of treatment and therefore improved anxiety outcomes. Patient Joviality has stronger statistical backing as a predictor in this model.

This model was estimated to explain 26.5% of the variation in the outcome scores. This compares to 20.4% of the variation estimated to be explained by the baseline model, suggesting a reasonable improvement with the inclusion of the linguistic predictors.

## 6.3 Overview of results

The results put forward in this chapter suggest some statistically significant associations between PANAS-X based linguistic features and PHQ-9 and GAD-7 scores reported both before and after a therapy session. Throughout the models it seems that patient joviality and therapist positive language were the most frequently recurring statistically significant predictors of outcome. These were both included in almost all models developed and were statistically significant in all those they were included in with one exception: therapist positive language in the prediction of end of treatment GAD-7 score. Therapist negative language was retained in a number of the models developed but with p-values around 0.1, there is only weak evidence behind the statistical association between this feature and outcome. These results suggest that the linguistic features of interest here are patient joviality and therapist positive language. Both aim to measure language within positive affect but the patient measure is narrower. As the broader measure of patient positive language was also tested in this chapter, this may suggest that in patient language, use of very upbeat language, is better associated with mental health outcome than more general positive affective language.

The nature of the relationship between patient joviality and outcome is more difficult to tease apart here as the association was statistically significant across all the models. It is therefore possible, and likely, that the relationship is bi-directional in that a better mental health outcome before a session is reflected in the language used in that session as well as being a reason for positivity in the session. Similarly, a treatment session carried out with more positive language is likely to improve future outcomes, both short and long term. Interestingly, in contrast with the results of the previous chapter, therapist positive language was more closely associated with outcome during the course of treatment as opposed to predictive of end of treatment outcome. In this case, positivity in therapist language could be a result of patient outcome measures and in turn, therapist positivity may improve patient short-term outcomes. The difference in results with the previous chapter does however suggest that the two measures are tapping in to slightly different types of positive language. This isn't surprising given that the LIWC categories are much broader but to investigate the specific differences would require further consideration of the terms included in both categories and the concepts they relate to.

# Chapter 7. Results from models fitted with Revised Cognitive Therapy Scale (CTS-R) based linguistic measures

This chapter presents the results of the mixed effects models developed with the I2E queries based on four Revised Cognitive Therapy Scale (CTS-R) items as linguistic predictor variables. These four candidate predictor variables were based on the following items: agenda setting, homework setting, pacing and interpersonal effectiveness. The development process for the queries extracting each of these linguistic features was described in the Methods chapter.

## 7.1 Description of the predictor variables

Four candidate predictors were tested in this section of analysis. Each of these was considered only in the therapist language as the scale was originally developed to focus on therapist skills and behaviour. The aim was to consider an association between linguistic evidence of the presence of these features of cognitive behaviour therapy in the session transcripts and outcome. As with previous linguistic features, the scores in Table 7-1 represent percentages of the language used by the therapist is each session that qualifies within each category.

**Table 7-1 Summary statistics of CTS-R based linguistic features**

| Linguistic feature | Mean score | SD | Min | Max |
|---|---|---|---|---|
| **Agenda setting (CTS-R)** | 0.06 | 0.13 | 0 | 1.36 |
| **Homework setting (CTS-R)** | 0.02 | 0.07 | 0 | 0.61 |
| **Pacing (CTS-R)** | 0.04 | 0.09 | 0 | 1.22 |
| **Interpersonal Effectiveness (CTS-R)** | 0.25 | 0.26 | 0 | 2.68 |

## 7.2 Model results

### 7.2.1 Outcome 1 – PHQ-9 score before session.

This model considers the association between the CTS-R based linguistic features, alongside baseline features, in a given therapy session for the PHQ-9 score recorded before that session. The model was fitted on a data set from 1758 appointments, from 374 individual patients.

**Table 7-2 Results from model predicting PHQ-9 score before session from CTS-R-based linguistic features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.70 | [ 0.63 ; 0.76 ] | <0.001 |
| ***Number of sessions*** | -0.43 | [ -0.49 ; -0.36 ] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.95 | [ 0.07 ; 1.90 ] | <0.001 |
| ***Step group 3+*** | 3.46 | [ 2.25 ; 4.67 ] | |
| ***Agenda setting*** | -1.83 | [ -3.22 ; 0.43 ] | 0.011 |
| ***Constant*** | 2.31 | [ 1.32 ; 3.31 ] | <0.001 |

The results (in Table 7-2) suggest that in addition to the baseline features previously described, one CTS-R based linguistic feature was statistically associated with outcome. This was the agenda setting feature. The proportion of language referring to agenda setting was associated with PHQ-9 score recorded before a treatment session with a coefficient of -1.8 (95%

CI [ -3.22 ; 0.43 ], *p* = 0.011). This suggests that for every percent more language used by the therapist that qualifies as agenda setting language in the query, the PHQ-9 score recorded before the therapy session was expected to be 1.8 points lower, on average. A higher proportion of references to agenda setting was therefore suggested to be associated with lower PHQ-9 scores and therefore a better depression outcome. Given that agenda setting would normally be instigated by the therapist, it may be that the severity of depression as suggested by the PHQ-9 score influences the frequency of references to agenda setting. This could be due to greater focus on the emotional experience with an individual with a higher depression score, for example. These results were maintained when the same predictors were tested using data only from individuals who had completed their course of treatment. The model was fitted on data from 1266 appointments involving 206 patients. The coefficient associated with agenda setting in this case was -1.72 (95% CI = [-3.36 ; -0.08], *p* = 0.04). Given the similarity between these two models, this second version will not be shown here but can be found in the appendix.

### 7.2.2 Outcome 2 – GAD-7 score before session.

This model considered the association between the CTS-R based linguistic features, alongside baseline features, in a given therapy session for the GAD-7 score recorded before that session. The model was fitted on a data set from 1741 appointments, from 370 individual patients.

The developed model suggests that two of the CTS-R based linguistic predictors were significantly associated with GAD-7 score reported just before a therapy session. These were interpersonal effectiveness and agenda setting. Both were negatively associated with outcome suggesting that higher proportions of language showing evidence of agenda setting and interpersonal effectiveness in the therapist’s language were associated with lower GAD-7 scores recorded before the session, suggesting a better anxiety outcome. The results for this model are presented in Table 7-3.

**Table 7-3 Results from model predicting GAD-7 score before a session from CTS-R-based linguistic features**

| *Predictors* | *B* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.62 | [ 0.54 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.74 | [ -0.94 ; -0.53 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.003 ; 0.03 ] | 0.018 |
| ***Diagnostic group (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.61 | [ -0.29 ; 1.51 ] | 0.112 |
| ***Diagnostic group 3 (Mixed)*** | 1.06 | [0.05 ; 2.06 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.08 | [ 0.09 ; 2.08 ] | <0.001 |
| ***Step group 3+*** | 3.47 | [ 2.21 ; 4.72 ] | |
| ***Interpersonal effectiveness*** | - 0.85 | [ -1.68 ; -0.03 ] | 0.041 |
| ***Agenda setting*** | -2.00 | [ -3.35 ; -0.65 ] | 0.004 |
| ***Constant*** | 3.32 | [ 2.04 ; 4.60 ] | <0.001 |

The equivalent model was fitted on a data set containing only data from individuals who completed their course of therapy. This was made up of data from 1250 appointments involving 202 patients. In this model neither interpersonal effectiveness nor agenda setting were significantly associated with outcome or reached the threshold for inclusion in the model. In this set of patients, none of CTS-R based language features were significantly associated with outcome. This suggests a marked difference in the association between these language features and outcome between the two populations, perhaps suggesting that agenda setting and interpersonal effectiveness do not have a strong impact on outcome when patients are engaged but do when they are not well engaged in treatment.

### 7.2.3 Outcome 3 – PHQ-9 score before next session.

This model considered the association between CTS-R based linguistic features at a given session and the PHQ-9 score taken just before the next session. The model was fitted on the full data set made up of data from 1685 sessions involving 372 patients and on the reduced data set of patients who completed treatment made up of 1196 sessions from 203 patients. None of the tested linguistic features were statistically significant in this model. When the model was tested on a dataset containing data from only those patients who completed their course of therapy, this result was the same. None of the CTS-R based linguistic features were associated with outcome.

### 7.2.4 Outcome 4 – GAD-7 score before next session.

This model considered the association between the CTS-R based linguistic features in a given therapy session and the GAD-7 score recorded before the next session. The model was fitted on a data set of 1683 appointments, from 369 individual patients.

The results from this model suggest that of the linguistic features tested, only the agenda setting feature was significantly associated with outcome. Agenda setting was associated with the GAD-7 score before the next session with a coefficient of -2.2 (95% CI = [ -3.65 : -0.72 ], $p$ = 0.003). This suggests that for every percent of the therapist's language that qualifies as agenda setting language as defined by the I2E query, the GAD-7 score taken just before the next session was expected to be 2.2 points lower, on average. This suggests that evidence of agenda setting in a therapy session is associated with improved short-term anxiety outcomes.

**Table 7-4 Results from model predicting GAD-7 score before next session from CTS-R-based linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.59 | [ 0.52 ; 0.67 ] | <0.001 |
| ***Number of sessions*** | -0.80 | [ -0.97 ; -0.61 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.01 ; 0.04 ] | 0.002 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.65 | [ -0.23 ; 1.54 ] | 0.099 |
| ***Diagnostic group 3 (Mixed)*** | 1.06 | [0.08 ; 2.04 ] | |
| ***Step group 2 –ref. group*** | | | |
| ***Step group 3*** | 1.23 | [ 0.25 ; 2.20 ] | <0.001 |
| ***Step group 3+*** | 3.64 | [ 2.41 ; 4.87 ] | |
| ***Agenda setting*** | -2.19 | [ -3.65 : -0.72 ] | 0.003 |
| ***Constant*** | 3.09 | [ 1.53 ; 3.94 ] | <0.001 |

When the equivalent model was developed using a data set from only individuals who completed their course of treatment made up of data from 1198 session from 200 patients, the homework setting feature was found to be statistically significant in addition to agenda setting (see Table 7-5). Homework setting was found to be negatively associated with outcome with a coefficient of -3.03 (95% CI = [ -5.94 ; -.13 ], *p*= 0.002) (compared to a non-significant association of -1.82 (95% CI = [-4.31 ; 0.68 ], *p* = 0.153) in the full data set) suggesting that for every percent of therapist language that qualifies as homework setting language in the I2E query, the GAD-7 score taken before the next session was expected to be 3.03 points lower, on average. Agenda setting was also negatively associated with outcome score, as was the case in the previously presented model. In the model using data from patients who completed treatment, the coefficient was -2.8 (95%CI = [-4.46 ; 1.19 ], *p*= 0.065), compared to -2.2 (95% CI = [ -3.65 : -0.72 ], *p* = 0.003) in the previous model, suggesting a stronger coefficient but a higher significance value making this less likely to be a true effect than in the case of the complete data set. These results suggest that more evidence of both agenda setting and homework setting in the therapist's language in a therapy session is associated with better anxiety scores reported before the next therapy session.

**Table 7-5 Results from model predicting GAD-7 score before next session from CTS-R-based features – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.55 | [ 0.44 ; 0.65 ] | <0.001 |
| ***Number of sessions*** | -0.85 | [ -1.06 ; -0.64 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.01 ; 0.04 ] | 0.009 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.13 | [ -0.02 ; 2.59 ] | 0.089 |
| ***Diagnostic group 3 (Mixed)*** | 1.28 | [0.08 ; 2.67 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.21 | [ -0.08 ; 2.50 ] | <0.001 |
| ***Step group 3+*** | 3.28 | [ 1.66 ; 4.90 ] | |
| ***Homework*** | -3.03 | [ -5.94 ; -.13 ] | 0.002 |
| ***Agenda setting*** | -2.82 | [-4.46 ; 1.19 ] | 0.065 |
| ***Constant*** | 3.10 | [ 1.55 ; 4.65 ] | <0.001 |

### 7.2.5 Cross-validation results

Table 7-6 presents the summary statistics from the cross-validation carried out on each of the models where CTS-R based linguistic features were statistically significant at the 15% level. These suggest that the stronger model, with the highest R-squared, is achieved when looking to predict the PHQ-9 score measured before a treatment session. The models looking at the two versions of GAD-7 as an outcome score were both weaker. When considering the cross-validated R-squared and calibration slope, the model predicting the GAD-7 score recorded before the next session did not seem weaker than that predicting the GAD-7 score reported before the current session. This stands in contrast to the pattern seen in previous results chapters where the models including the time lag were consistently weaker than the cross-sectional models.

**Table 7-6 Summary results from five-fold cross-validation**

| Outcome | All or completed cases | Mean cross-validated $R^2$ | Range of $R^2$ | Calibration slope | Intercept |
|---|---|---|---|---|---|
| **Outcome 1 – PHQ-9** | All cases | 0.51 | [ 0.40 – 0.65] | 0.98 | 0.20 |
| **before session** | Completed only | 0.43 | [ 0.34 – 0.62] | 0.92 | 0.40 |
| **Outcome 2 – GAD-7** | All cases | 0.36 | [0.27 – 0.54] | 0.93 | 0.67 |
| **before session** | Completed only | - | - | - | - |
| **Outcome 3 – PHQ-9** | All cases | - | - | - | - |
| **before next session** | Completed only | - | - | - | - |
| **Outcome 4 – GAD-7 score** | All cases | 0.35 | [ 0.21 – 0.53] | 0.91 | 0.93 |
| **before next session** | Completed only | 0.32 | [ 0.19 – 0.48] | 0.92 | 0.98 |

### 7.2.6 Outcome 5 – Final PHQ-9 score

This model considered the mean levels of CTS-R based linguistic features during the first two treatment sessions and their association with the final PHQ-9 score reported prior to the final therapy appointment. The four CTS-R based measures were tested in a model containing the previously statistically significant baseline PHQ-9 score. This model was fitted on data from 207 patient cases.

**Table 7-7 Results of linear regression predicting final PHQ-9 score from baseline features and CTS-R-based linguistic features early in treatment**

| ***Final PHQ-9 score*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.49 | [ 0.38 ; 0.59 ] | <0.001 |
| ***Pacing language (CTS-R based I2E query)*** | 8.40 | [ -2.23 ; 19.03] | 0.121 |
| ***Constant*** | 0.89 | [ -0.55 ; 2.34 ] | 0.226 |

The results suggest that only the measure of pacing language (CTS-R) was retained in this model with a coefficient of 8.4 (95% CI = [-2.23 ; 19.03], *p* = 0.121). This high coefficient, the wide confidence interval, and the high p-value indicate caution on interpreting these results. The CTS-R language features were not very frequent in the transcripts, meaning that proportion scores were generally low. The coefficient is expressed in terms of the effect of a one unit, in this case percentage, change in the pacing variable value.

The variation in outcome scores explained by this model is estimated to be 29.6%, which is only a limited (less than 1%) improvement on the baseline model.

### 7.2.7 Outcome 6 – Final GAD-7 score

This model considered the mean levels of CTS-R based linguistic features during the first two treatment sessions and their association with the final GAD-7 score reported prior to the final therapy appointment. The four CTS-R based measured were tested in a model accounting for the baseline PHQ-9 score. This model was fitted on data from 203 patient cases. No table with summary statistics is presented here as none of the CTS-R features included in analysis were found to be significantly associated with outcome in this model.

## 7.3 Overview of results

The results presented in this chapter suggest that, of the four CTS-R based linguistic features presented, agenda setting was the only one that recurred as statistically significant at the 15% level in a number of models. These were of both outcomes before a session and of GAD-7 score before the next session. This is interesting as it is a therapist language measure taken after the recorded outcome. This suggests that the level of depression and anxiety as recorded by the outcomes may influence the level of agenda setting references in the session. In the model of GAD-7 score before the next session, the association suggests that a higher number of references to agenda setting, and therefore close adherence to a structured CBT session, may improve short-term outcomes. However, this is speculation and the nature of this relationship is still unclear.

The other three features made an appearance in one model each but did not appear to be consistently associated with outcome scores. Additionally, the cross-validation results suggest that only small gains were made with the inclusion of the CTS-R variables when one or more of these were predictive. In a number of cases, however, none of the variables were statistically significant. The results associated with these language features therefore were not compelling. This may be to do with the low values within each linguistic feature, a sign that these elements of language are either not being picked up adequately within the transcripts, or that they are only rare features within these transcripts. Sensitivity analyses would provide some insight into this. It may be more important to focus initially on correct identification of these features and determining their presence in therapy transcripts before going to the next step and looking at their association with outcome. This is an idea that will be further discussed in later chapters.

# Chapter 8. Results from combined models

This chapter presents the results of combined models which incorporated variables retained in the models in previous analyses as candidate predictor variables. The same six outcome variables were considered as in previous chapters and for each model developed, the statistically significant variables from each individual set of linguistic features were entered into the model. The model was developed following the same procedure of backwards stepwise variable selection using a significance level of 0.15 as a threshold for inclusion in the model.

## 8.1 Model results

### 8.1.1 Outcome 1 – PHQ-9 score just before the session

This model considers a set of linguistic features that have previously shown an association with the closest PHQ-9 score. The model was developed to look at how these linguistic features, measured during a given therapy session, were associated with the PHQ-9 score reported just before that session. The model was fitted on data from 1758 appointments from 374 individual patients. The results from this model can be found in Table 8-1.

**Table 8-1 Results from model predicting PHQ-9 score before session from combined linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.33 | [ -0.39 ; -0.26 ] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.56 | [ -0.36 ; 1.48 ] | <0.001 |
| ***Step group 3+*** | 2.91 | [ 1.75 ; 4.08 ] | |
| ***Patient Negative language (LIWC)*** | 0.28 | [ 0.10 ; 0.45 ] | 0.002 |
| ***Patient Positive language (LIWC-based I2E query)*** | -0.29 | [ -0.54 ; -0.04 ] | 0.025 |
| ***Patient Negations (LIWC)*** | 0.22 | [ 0.02 ; 0.42 ] | 0.033 |
| ***Patient Social language (LIWC)*** | 0.36 | [ 0.12 ; 0.61 ] | 0.004 |
| ***Patient First person plural pronouns*** | -0.34 | [ -0.76 ; 0.08 ] | 0.114 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.31 | [ -0.64 ; 0.03 ] | 0.079 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.49 | [ -0.80; -0.18 ] | 0.002 |
| ***Therapist certainty language (LIWC)*** | -1.05 | [ -2.39 ; 0.28 ] | 0.123 |
| ***Agenda setting*** | -1.83 | [ -3.20 ; -0.45 ] | 0.009 |
| ***Constant*** | 2.73 | [ 1.36 ; 4.09 ] | <0.001 |

The results of the model suggest that the baseline predictors that were significantly associated with outcome also were after the inclusion of the combined linguistic predictors in this model. In addition to these, nine linguistic predictors were retained in this model. These were six patient language features and three therapist language features. Of the patient language features, four were from the set of LIWC language features. These were patient negative language (LIWC), patient use of negations, patient social language and patient use of first person plural pronouns. Patient use of first person plural pronouns was suggested to be negatively associated with outcome and the other three patient LIWC features were positively associated with outcome suggesting that higher levels of these features were associated with a higher PHQ-9 score recorded before the session. Individual coefficients and significance values can be found in Table 8-1. The two remaining patient language features that were included in this model were patient positive language (LIWC-based I2E query) and joviality (Expanded PANAS-X category). Patient positive language (LIWC-based I2E query) was negatively associated with outcome with a coefficient of -0.29 (95% CI = [ -0.54 ; -0.04 ] $p$ = 0.025), suggesting that for every percent of a patient's language in a therapy session that fits within the I2E positive query, the PHQ-9 score before the session was expected to be an average of 0.29 points lower. Patient joviality (expanded PANAS-X category) was also negatively associated with outcome, with only a slightly larger coefficient of -0.31 (95% CI = [ -0.64 ; 0.03 ], $p$ = 0.079).

The three other language features in this model were found within therapist language. All three of these were suggested to be negatively associated with outcome, suggesting that higher levels of these were associated with lower outcome score, and thus improved depression outcomes. Therapist positive language as measured by LIWC-based I2E query was statistically significant with a coefficient of -0.49 (95% CI = [ -0.80; -0.18 ], $p$ = 0.002). The final predictor variable significantly associated with outcome in this model was that measuring references to agenda setting. Agenda setting was associated with PHQ-9 score reported before the session with a coefficient of -1.8 (95% CI = [ -3.20 ; -0.45 ], $p$ = 0.009) suggesting that for every percent of therapist language that refered to agenda setting, the PHQ-9 score before the session was expected to be 1.8 points lower, on average.

The majority of the variables retained in these models were suggested to be associated with outcome with low attached p-values, suggesting reasonably strong evidence of the presence of an effect. However, patient use of first person plural language and therapist certainty language both had high attached p-values, suggesting that there is only weak evidence supporting their associations with outcome.

When the equivalent model was considered in a dataset containing data from only individuals who completed their course of treatment, a number of differences were apparent. It was fitted on a set of data containing information from 1266 appointments from 206 individuals. The fixed results are presented in Table 8-2.

**Table 8-2 Results from model predicting PHQ-9 score before session from combined linguistic features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.62 | [ 0.54 ; 0.70 ] | <0.001 |
| ***Number of sessions*** | -0.38 | [ -0.46 ; -0.31 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.20 | [ -0.006 ; 2.41 ] | <0.001 |
| ***Step group 3+*** | 3.44 | [ 1.91 ; 4.97 ] | |
| ***Patient Negative language (LIWC)*** | 0.41 | [ 0.20 ; 0.63 ] | <0.001 |
| ***Patient Positive language (LIWC-based I2E query)*** | -0.32 | [ -0.57 ; -0.06 ] | 0.020 |
| ***Therapist Negations (LIWC)*** | 0.34 | [ -0.05 ; 0.74 ] | 0.088 |
| ***Patient Social language (LIWC)*** | 0.44 | [ 0.14 ; 0.75 ] | 0.005 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.43 | [ -0.80; -0.07 ] | 0.020 |
| ***Agenda setting*** | -1.74 | [ -3.36 ; -0.12 ] | 0.035 |
| ***Constant*** | 1.91 | [ 0.26 ; 3.55 ] | 0.023 |

The results from this model suggest that three patient language features and three therapist language features were retained in this model. Of the patient language features, two were categories from the LIWC dictionary; these were patient negative language (LIWC) and patient social language. Both were positively associated with outcome, suggesting that higher levels of patient negative language and patient social language as measured by the LIWC were associated with higher PHQ-9 scores before the session. The third patient language feature statistically and significantly associated with outcome in this model was the LIWC-based I2E query measure of patient positive language. Patient positive language (LIWC-based I2E query) was statistically significantly associated with PHQ-9 score with a coefficient of -0.32 (95% CI = [ -0.57 ; -0.06 ] ,$p$ = 0.014) suggesting that for every percent of patient language in a therapy session that qualified as positive by the I2E query definition, the PHQ-9 score before the session was expected to be 0.32 points lower, on average. More patient positive language was therefore associated with better a depression outcome.

Three features within therapist language were included in this model. Therapist use of negations (LIWC category) was positively associated with outcome with a coefficient of 0.34 (95% CI = [ -0.05 ; 0.74 ] ,$p$ = 0.088). This suggests that for every percent of therapist language in a session that fits within the negations LIWC category, the PHQ-9 score before the session was expected to be 0.34 points higher, on average. However, the higher p-value associated with this variables suggests the evidence supporting the association is weaker than for the other variables in this model. Interpretable in the same way but with an opposite direction of association were the final two language features that were statistically significant in this model at the 15% level: therapist positive language as measured by the I2E query based on the expanded PANAS-X and agenda setting language. Both were negatively and significantly associated with outcome suggesting that higher levels of positive language (expanded PANAS-X-based query) and agenda setting language were associated with a lower PHQ-9 score and therefore a better depression outcome.

### 8.1.2 Outcome 2 – GAD-7 score just before the session.

This model considered the predictors retained in models in previous chapters as candidate predictors when these were combined. It looked at the association between these candidate predictors and the GAD-7 score that the patient is requested to report before a therapy session. The model was fitted on data from 1741 appointments from 370 individual patients. The results from this model can be found in Table 8-3.

**Table 8-3 Results from model predicting PHQ-9 score before session from combined linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.60 | [ 0.53 ; 0.67] | <0.001 |
| ***Number of sessions*** | -0.61 | [ -0.81 ; -0.41 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.01 | [ 0.001 ; 0.03] | 0.046 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.68 | [ -0.18 ;1.5] | 0.008 |
| ***Diagnostic group 3 (Mixed)*** | 1.09 | [ 0.12 ;2.06] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 0.74 | [ -0.23 ; 1.70] | <0.001 |
| ***Step group 3*** | 2.97 | [ 1.76 ; 4.19 ] | |
| ***Patient Negative language (LIWC)*** | 0.23 | [ 0.06 ;0 .41 ] | 0.009 |
| ***Patient Negations (LIWC)*** | 0.15 | [ -0.04 ;0.34 ] | 0.112 |
| ***Patient Social language (LIWC)*** | 0.36 | [ 0.12 ; 0.59 ] | 0.003 |
| ***Patient first person plural pronouns (LIWC)*** | -0.37 | [ -0.77 ; 0.03 ] | 0.070 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.41 | [ -0.71 ; -0.12 ] | 0.007 |
| ***Therapist negative language (LIWC-based I2E query)*** | 0.26 | [ 0.06 ; 0.51 ] | 0.045 |
| ***Therapist certainty (LIWC)*** | -1.21 | [ -2.50 ; -0.07 ] | 0.064 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.40 | [ -0.67 ; -0.13 ] | 0.004 |
| ***Agenda setting*** | -1.93 | [ -3.27 ; -0.59 ] | 0.005 |
| ***Constant*** | 2.81 | [ 1.28 ; 4.35 ] | <0.001 |

Beyond the baseline predictors, nine linguistic features, from the four different sets of variables, were retained in this model. Compared to the equivalent model predicting PHQ-9 score reported in Table 8-1, there were two differences in the set of predictors included in the final model. In this model, patient positive language (LIWC-based I2E query) was not statistically significantly associated with outcome, whereas it had been previously. However, therapist negative language as measured by the LIWC-based query was associated with outcome in this model with a positive coefficient of 0.26 (95% CI = [ 0.006 ; 0.51 ], *p* = 0.045). This suggests that a higher proportion of therapist negative language as measured by the I2E query was associated with higher GAD-7 score before the therapy session and therefore worse anxiety outcomes. The remaining associations between linguistic features and outcome score reported in the equivalent model predicting PHQ-9 score above were also included here with coefficients of similar magnitude and the same direction of association.

When the equivalent model was developed on a data set containing only the data from individuals who had completed their course of treatment, a few differences appeared. The results of this model development can be found in Table 8-4. The model was fitted on data from 1250 appointments from 202 individual patients. Patient negative language (LIWC), patient social language (LIWC) and patient joviality (expanded PANAS-X) were all associated with outcome in this model with the same direction of association as in the previously described model and with slightly stronger coefficients. Patient use of first person plural pronouns and patient use of negations were not retained in this model as they were in the model above. In terms of therapist language features, some differences were also apparent. Therapist negative language as measured by the LIWC-based I2E query was not significantly associated with outcome as it had been in the previous model. Therapist insight language (LIWC) was suggested to be associated with outcome in this model when it had not been previously, but the associated significance value of 0.148 put this predictor on the very edge of inclusion in the model and suggests the evidence supporting the reality of this association is weak. Both therapist positive language as measured by the PANAS-X based I2E query and agenda setting were negatively associated with the GAD-7 score reported before the therapy session, as was the case in the previous models.

**Table 8-4 Results from model predicting GAD-7 score before a session from combined linguistic features – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.56 | [ 0.46 ; 0.66] | <0.001 |
| ***Number of sessions*** | -0.67 | [ -0.91 ; -0.44 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.01 | [ -0.002 ; 0.03] | 0.084 |
| ***Diagnostic group 1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.21 | [ -0.07 ;2.36] | 0.078 |
| ***Diagnostic group 3 (Mixed)*** | 1.20 | [ -0.08 ;2.49] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.80 | [ -0.46 ; 2.07] | 0.002 |
| ***Step group 3+*** | 1.21 | [ 1.17 ; 4.37 ] | |
| ***Patient Negative language (LIWC)*** | 0.32 | [ 0.12 ;0 .53 ] | 0.002 |
| ***Patient Social language (LIWC)*** | 0.33 | [ 0.12 ; 0.59 ] | 0.027 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.59 | [ -0.93 ; -0.25 ] | 0.001 |
| ***Therapist insight (LIWC)*** | 0.13 | [ -0.04 ; 0.30 ] | 0.148 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.41 | [ -0.73 ; -0.08 ] | 0.015 |
| ***Agenda setting*** | -1.98 | [ -3.54 ; -0.43 ] | 0.013 |
| ***Constant*** | 2.77 | [ 0.83 ; 4.72 ] | <0.001 |

### 8.1.3 Outcome 3 – PHQ-9 score before the next session

In this model, the linguistic features at one therapy session were considered as predictors of the PHQ-9 score measured before the next therapy session. The model was fitted on data from 1685 appointments from 372 individual patients. The results for this model can be found in Table 8-5.

**Table 8-5 Results from model predicting PHQ-9 score before next session from combined linguistic features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.68 | [ -0.88 ; -0.48 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | { 0.01 ; 0.04 ] | 0.004 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.80 | [ - 0.14 ; 1.75 ] | <0.001 |
| ***Step group 3*** | 3.41 | [ 2.23 ; 4.60 ] | |
| ***Therapist Negations (LIWC)*** | 0.44 | [ 0.12 ; 0.96 ] | 0.015 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.40 | [ -0.69 ; 0.11 ] | 0.008 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.44 | [ -0.79 ; -0.09 ] | 0.013 |
| ***Constant*** | 3.01 | [ 2.55 ; 5.79] | <0.001 |

Three of the language features tested were retained in the model and suggested to be statistically significant at the 15% level. These were therapist use of negations (LIWC), therapist use of positive language as measured by the I2E query based on the expanded PANAS-X positive category and patient joviality as measured by the expanded PANAS-X category. Therapist negation use was positively associated with outcome with a coefficient of 0.44 (95% CI =[ 0.12 ; 0.96 ], *p* = 0.015), suggesting that for every percent of therapist language that fits within the negation LIWC category, the PHQ-9 score before the next therapy session was expected to be 0.44 of a point higher, on average. Therapist positive language (Expanded PANAS-X based query) and patient joviality (Expanded PANAS-X category) were both negatively associated with outcome suggesting that higher levels of these linguistic features within a therapy session were associated with lower PHQ-9 scores before the next session and therefore an improved depression outcome.

When the equivalent model was developed on a dataset containing only data from those individuals who completed their course of treatment (1196 appointments from 203 individual patients) there was only one notable difference. The three linguistic features that were statistically significant in the previous model were also statistically significant here in the same direction of association and with slightly larger coefficient sizes. These were patient joviality expanded PANAS-X), therapist use of negations and therapist positive language as measure by the expanded PANAS-X-based I2E query. In addition to these, one further linguistic feature was included in the model, and with an attached p-value on the border of the lower significance threshold. This was patient use of first person singular pronouns (‘I’). This feature was negatively associated with outcome ($b$ = -0.23, 95% CI = [ -0.46 ; 0.01 ], $p$ = 0.055), suggesting that higher levels of first person singular pronouns were associated with a lower PHQ-9 score before the next therapy session. This suggests, for example, that a patient who used 2% first person singular pronouns in their session was likely to provide a PHQ-9 score prior to the next session that was 0.23 points lower, on average, than an individual who used 1% positive language. The results for this model can be found in Table 8-6.

**Table 8-6 Results from model predicting PHQ-9 score before next session from combined linguistic features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.61 | [ 0.53 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.71 | [ -0.96 ; -0.48 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | { 0.003 ; 0.04 ] | 0.021 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.32 | [ 0.07 ; 2.56 ] | <0.001 |
| ***Step group 3+*** | 3.71 | [ 2.16 ; 5.25 ] | |
| ***Therapist Negations (LIWC)*** | 0.54 | [ 0.12 ; 0.96 ] | 0.012 |
| ***Patient First person singular pronouns*** | -0.23 | [ -0.46 ; 0.01 ] | 0.055 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.51 | [ -0.86 ; 0.14 ] | 0.006 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.46 | [ -0.87 ; -0.04 ] | 0.030 |
| ***Constant*** | 4.17 | [ 2.55 ; 5.79] | <0.001 |

### 8.1.4 Outcome 4 – GAD-7 score before the next session.

This model considered the predictor variables that were retained in the model in previous analyses within a combined model looking to use linguistic features from a given therapy session to predict GAD-7 reported before the next therapy session. This model was fitted on data from 1683 therapy sessions involving 369 patients. The results for this model can be found in Table 8-7

**Table 8-7 Results from model predicting GAD-7 score before next session from combined linguistic features**

| *Fixed-effects* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.59 | [ 0.51 ; 0.67] | <0.001 |
| ***Number of sessions*** | -0.73 | [ -0.91 ; -0.54 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.01 ; 0.04] | 0.003 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.74 | [ -0.14 ;1.61 ] | 0.072 |
| ***Diagnostic group 3 (Mixed)*** | 1.09 | [ 0.13 ;2.06] | |
| ***Step group 2*** | 1.09 | [ 0.12 ; 2.05 ] | <0.001 |
| ***Step group 3*** | 3.53 | [ 2.32 ; 4.74 ] | |
| ***Therapist Negative language (LIWC)*** | 0.26 | [ 0.08 ; 0.44 ] | 0.004 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.57 | [ -0.90 ; -0.25 ] | 0.001 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.28 | [ -0.56 ; -0.004 ] | 0.046 |
| ***Agenda setting*** | -2.01 | [ -3.46 ; -0.55 ] | 0.007 |
| ***Constant*** | 3.04 | [ 1.66 ; 4.42 ] | <0.001 |

The results of this model suggest that four linguistic predictors were retained in the model and significantly associated with outcome. Of these, only one related to patient language whereas the other three were features found in the language used by the therapist. Patient joviality (expanded PANAS-X) was negatively associated with outcome with a coefficient of -0.28 (95% CI = [ -0.56 ; -0.004 ], $p$ = 0.046) suggesting that for every percent of patient language in a given therapy session that fit within the expanded PANAS-X joviality category, the GAD-7 score reported before the next therapy session was likely to be 0.28 points lower, on average. Two different measures of affect in therapist language were also significantly associated with outcome in the model. Therapist negative language as measured by the LIWC dictionary was positively associated with GAD-7 score measured before the next session. This suggests that higher levels of therapist negative language (LIWC) were associated with a worse anxiety outcome recorded before the next treatment session. Therapist positive language was also associated with outcome as had been the case in the PHQ-9 version of the model and with a slightly larger coefficient of -0.58 (95% CI = [-0.90 ; -0.25], *p* = 0.001) as compared to -0.44 (95% CI = [-0.79 ; -0.09 ], *p*= 0.013) previously. The final linguistic feature in this model was agenda setting, which was negatively associated with outcome. This suggests that a higher proportion of references to agenda setting within a session was associated with a lower GAD-7 score reported before the next therapy session.

When the equivalent model was fitted on a set of data from only patients who completed their course of therapy, a number of additional predictors were retained in the model (see Table 8-8). This model was fitted on data from 1198 appointments involving 200 patients. The four predictors described in the previous model were included in this model, with the addition of patient positive and negative language based on the expanded PANAS-X, and homework setting. The expanded PANAS-X based measure of patient positive language was negatively associated with outcome with a coefficient of -0.33 (95% CI = [ -0.72 ; 0.06 ] ,*p* = 0.098) and the expanded PANAS-X based measure of patient negative language was positively associated with outcome with a coefficient of 0.41 (95% CI =[ 0.01 ; 0.81 ] ,*p* =0.046). These were suggested to affect the GAD-7 score before the next session in opposing directions, with higher levels of positive language associated with a lower GAD-7 before the next session and the opposite being the case for negative language. The third additional linguistic feature that was associated with outcome in this model as compared to the previous model was homework setting, which was negatively associated with outcome with a coefficient of -3.34 (95% CI = [ -6.22 ; -0.46 ], *p* = 0.023). This suggests that for every percent of therapist language in a session that fits within the query definition of homework setting language, the GAD-7 score before the next session was expected to be 3.34 points lower.

It is also important to note that the significance values associated with Patient Joviality and patient positive language are both around 0.1, suggesting only moderate to weak evidence of these two predictors being associated with outcome.

**Table 8-8 Results from model predicting GAD-7 score before next session from combined linguistic features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.53 | [ 0.44 ; 0.63] | <0.001 |
| ***Number of sessions*** | -0.77 | [ -0.99 ; -0.56 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ -0.004 ; 0.04] | 0.013 |
| ***Diagnostic group 1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.18 | [ 0.06 ;2.31] | 0.075 |
| ***Diagnostic group 3 (Mixed)*** | 1.23 | [ -0.03 ;2.51] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.06 | [ -0.20 ; 2.31] | <0.001 |
| ***Step group 3+*** | 3.03 | [ 1.46 ; 4.61 ] | |
| ***Therapist Negative language (LIWC)*** | 0.21 | [ -0.004 ; 0.42 ] | 0.055 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.52 | [ -0.91 ; -0.13 ] | 0.009 |
| ***Patient positive language (Expanded PANAS-X based I2E query)*** | -0.33 | [ -0.72 ; 0.06 ] | 0.098 |
| ***Patient negative language (Expanded PANAS-X based I2E query)*** | 0.41 | [ 0.01 ; 0.81 ] | 0.046 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.29 | [ -0.64 ; -0.06 ] | 0.104 |
| ***Agenda setting*** | -2.73 | [ -4.35 ; -1.410 ] | 0.001 |
| ***Homework*** | -3.34 | [ -6.22 ; -0.46 ] | 0.001 |
| ***Constant*** | 3.58 | [ 1.81 ; 5.35 ] | <0.001 |

### 8.1.5 Cross-validation of mixed effects models

The cross-validation results (see Table 8-9) suggest that each of the models developed provides a reasonable prediction of the different outcome scores. As was the case in previous chapters, the model predicting the PHQ-9 score before a therapy session using data from the same session had the strongest associated mean R-squared value of 0.54. The model predicting PHQ-9 score before the next session was slightly weaker with a mean cross-validated R-squared of 0.49. When the equivalent models were developed in a data set containing only data from completed patients, the models were again a little weaker but still reasonably strong. In the case of models with GAD-7 score as an outcome, these explained approximately 10% less of the variation in outcome scores than their PHQ-9 score equivalent, with the same pattern of small differences between the model versions. Overall, it appears that the addition of linguistic features to the baseline features in these models adds between 2 and 5% to the mean R-squared and therefore to the estimated variation in the data explained.

**Table 8-9 Summary results from five-fold cross-validation**

| Outcome | All or completed cases | Mean cross-validated $R^2$ | Range of $R^2$ | Calibration slope | Intercept |
|---|---|---|---|---|---|
| **Outcome 1 – PHQ-9 before session** | All cases | 0.54 | [ 0.42 – 0.66] | 0.99 | 0.11 |
| | Completed only | 0.47 | [ 0.39 – 0.64] | 0.97 | 0.07 |
| **Outcome 2 – GAD-7 before session** | All cases | 0.40 | [0.29 – 0.55] | 0.96 | 0.40 |
| | Completed only | 0.36 | [ 0.26 – 47] | 0.96 | 0.45 |
| **Outcome 3 – PHQ-9 before next session** | All cases | 0.49 | [0.38 – 0.62] | 0.96 | 0.47 |
| | Completed only | 0.41 | [ 0.30 – 0. 61] | 0.95 | 0.64 |
| **Outcome 4 – GAD-7 score before next session** | All cases | 0.36 | [ 0.23 – 0.53] | 0.92 | 0.89 |
| | Completed only | 0.34 | [ 0.19 – 0.47] | 0.92 | 0.88 |

### 8.1.6 Outcome 5 – Final PHQ-9 score

This model considered the statistically significant predictors at the 15% level of final outcome score within the individual sets of language features and case and baseline information. As previously, the linear regression model was developed using only data from individuals who had completed their course of therapy and considered levels of language features in the first two treatment sessions and their association with PHQ-9 score at the last session. This model was developed on data from 207 patients. The linear regression results can be found in Table 8-10.

**Table 8-10 Results of linear regression predicting final PHQ-9 score from baseline features and combined linguistic features early in treatment**

| *Final PHQ-9 score* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.46 | [ 0.35 ; 0.56 ] | <0.001 |
| ***Patient Positive language (Expanded PANAS-X based I2E query)*** | -2.24 | [ -3.88 ; -0.60 ] | 0.008 |
| ***Patient Negations (LIWC)*** | 0.66 | [ -0.24 ; 1.56 ] | 0.147 |
| ***Patient Social language (LIWC)*** | 0.98 | [ 0.12 ; 1.95 ] | 0.047 |
| ***Therapist positive language (LIWC-based I2E query)*** | -1.41 | [ -2.34 ; -0.48 ] | 0.003 |
| ***Constant*** | 4.08 | [ 0.48 ; 7.69 ] | 0.027 |

The results suggest that four linguistic features were retained in this model in addition to the baseline PHQ-9 score reported before the assessment session. These were patient social language (LIWC), patient use of negations (LIWC), therapist positive language (LIWC-based I2E query) and patient positive language (expanded PANAS-X based I2E query). In each case, the features considered refer to their mean levels in the first two treatment sessions as predictors of PHQ-9 score at the final session. Patient social language and patient use of negations were positively associated with outcome. The coefficient associated with social language use was 0.98 (95% CI = [ 0.12 ; 1.95 ], $p$ = 0.047). This suggests that if mean social language use was one percent higher during the first two treatment sessions, this was associated with an average of a 0.98 point higher PHQ-9 score reported at the final therapy session. Higher mean negation use in the first two treatment sessions was also associated with higher final PHQ-9 score but with a lower coefficient of 0.66 (95% CI = [ -0.24 ; 1.56 ], *p* = 0.147). However, the attached p-value of 0.147 is only just below the threshold for inclusion in the model and it suggests very weak evidence supporting the association. The two remaining linguistic predictors that were included in this model were suggested to be negatively associated with the final outcome score suggesting that higher levels of therapist positive language (LIWC-based query) and patient positive language (PANAS-X based query) were associated with a lower end of treatment PHQ-9 score, and therefore improved depression outcomes. For example, the coefficient associated with therapist positive language (LIWC-based I2E query) was -1.41 (95% CI = [ -2.34 ; -0.48 ], *p* = 0.003) suggesting that a 1% higher proportion of mean positive language use from a therapist during the first two treatment sessions was associated with a 1.41 point lower PHQ-9 score reported prior to the final therapy session, on average.

This model was suggested to explain 37% of the variation in the outcomes, suggesting a reasonable model overall. The baseline model was suggested to explain 28% of the variation in the outcome scores suggesting that there is some small gain from the addition of linguistic features.

### 8.1.7 Outcome 6 – Final GAD-7 score

This linear regression model was developed using only data from individuals who had completed their course of therapy and considered levels of language features in the first two treatment sessions and their association with GAD-7 score at the last session. The model was developed on data from 203 individuals. The linear regression results for this analysis can be found in Table 8-11.

**Table 8-11 Results of linear regression predicting final GAD-7 score from baseline features and combined linguistic features early in treatment**

| ***Final GAD-7 score*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.40 | [ 0.29 ; 0.51 ] | <0.001 |
| ***Patient Positive language (LIWC)*** | -0.72 | [ -1.18 ; -0.26 ] | 0.002 |
| ***Patient Negations (LIWC)*** | 1.12 | [ 0.32 ; 1.91 ] | 0.006 |
| ***Patient Social language (LIWC)*** | 1.18 | [ 0.31 ; 2.04 ] | 0.008 |
| ***Therapist positive language (LIWC-based I2E query)*** | -1.03 | [ -1.86 ; -0.21 ] | 0.015 |
| ***Therapist Negations (LIWC)*** | -1.24 | [ -2.47 ; -0.01 ] | 0.048 |
| ***Constant*** | 3.06 | [ -0.54 ; 6.66 ] | 0.095 |

The results suggest that mean levels of five linguistic features early in therapy were included in the model of GAD-7 score reported before the last therapy session. Of these, three were features of patient language and two were features of therapist language. Within the patient language features were three LIWC categories: positive language, use of negations and social language. Patient use of positive language (LIWC) was negatively associated with final GAD-7 score suggesting that a higher mean level of positive language in patient language early in therapy was associated with a lower GAD-7 score at the end of treatment. Patient negation use (LIWC) and patient social language (LIWC) were positively associated with outcome, suggesting that higher mean levels of these language features in patient language early in therapy were associated with a higher GAD-7 score at the end of treatment. Two therapist language features were significantly associated with outcome. Therapist positive language (LIWC-based I2E query) was negatively associated with outcome with a coefficient of -1.03 (95% CI = [ -1.86 ; -0.21 ] , $p$ = 0.015) suggesting that a 1% higher mean level of therapist positive language in the first two treatment sessions was associated with a 1.03 points lower, on average, GAD-7 score at the end of treatment. Therapist use of negations (LIWC) was also negatively associated with outcome, but with a coefficient of -1.24 (95% CI = [ -2.47 ; -0.01 ] ,$p$ =

0.115), suggesting that higher levels of negations in therapist language early in treatment were associated with a lower GAD-7 score at the end of treatment. This stands in contrast to the effect associated with patient use of negations, which were positively associated with outcome.

This combined model was estimated to explain 33% of the variation in the outcome scores in the data used for analysis. This is a reasonably strong model and the value of including linguistic predictors can be seen when comparing with a baseline model. The baseline model was estimated to explain 20% of the variation in the data.

## 8.2 Overview of results

The results from this chapter suggest that a range of linguistic features from the different sets developed were statistically significant predictors within the models presented at the 15% level. Throughout the models, measures of sentiment appeared to feature strongly, whether these were measured through the LIWC dictionary or text mining queries based on the LIWC or PANAS-X. The range of predictors that were retained was broader in the models of outcome reported just before a session. Models predicting outcome before the next session appeared to more closely revolve around negative and positive language measures. The smaller set of statistically significant predictors may be associated with the distance from reported outcome scores as it is a more difficult task than predicting an outcome score closer in time.

The information that these results can provide about the nature of these associations is limited but comparison across the different models developed can provide some clues towards this. For example, patient negative language (LIWC) was statistically significantly associated with outcome recorded before a session but not outcome before the next session. This could point towards negative language as reflecting mental health state and put it forward as a potential marker of this in language. However, the level of negative language in a treatment session does not appear to be a good indicator of short or long-term outcomes. In contrast, patient joviality may be both reflective of patient mental state and an indicator of short-term outcome. In this case there may be an immediate effect of positivity in improving patient measures of outcome. This does not appear to extend to long term outcomes however, as patient joviality is not statistically significant in the models of end of treatment outcome.

Across the range of models, cross-validation results suggest reasonably good calibration of the models in this dataset and some additional explained variation attributable to these models. The effects of the linguistic features appear to be statistically significant but the magnitude of their impact on the mixed effects models may be limited when considering the application of these models in practice. The improvement in the models predicting outcome at the end of treatment, however, is far greater. The inclusion of the linguistic features added between 9 and 13% in absolute terms in the variation in outcome explained. This increased overall R-squared values for the model by approximately 50% relative to the baseline model suggesting they may be useful predictors of end of treatment outcome. The external validation of these models on a dataset from a different population will provide further information on the value of the predictive models presented.

# Chapter 9. Results from external validation of outcome prediction models

This chapter presents the results of the external validation of models presented in Chapter 9. The previously developed models were tested on a new data set, called the 'validation data set'. For each outcome, the parameters were estimated on the development data set and predicted outcome values were then calculated from these. R-squared and calibration slope values were estimated and graphical representations of predicted and observed values as well as residual values were developed to assess model performance.

## 9.1 Descriptive statistics

A selection of descriptive statistics that are relevant to the models presented are included below. The distribution of step group and summary statistics for baseline PHQ-9 and GAD-7 scores in the validation set are presented in Table 9-1 and Table 9-2. Further details including age and provisional diagnoses of patients included in this data set were presented in Chapter 3 (Methods).

**Table 9-1 Step group frequencies**

| Step | Frequency | Per cent |
|---|---|---|
| **Assessment** | 26 | 6.9 |
| **Step 2** | 76 | 20.2 |
| **Step 3** | 251 | 66.7 |
| **Step 3+** | 23 | 6.1 |
| **Total** | 376 | 100.0 |

**Table 9-2 Summary statistics of baseline outcome scores**

| Outcome score | Mean | SD | Min | Max |
|---|---|---|---|---|
| **Baseline PHQ-9 score** | 12.22 | 6.63 | 0 | 27 |
| **Baseline GAD-7 score** | 11.54 | 6.14 | 0 | 21 |

The mean baseline values in the development data set were similar to those presented here with a mean baseline PHQ-9 of 12.60 (SD=6.46) and a mean baseline GAD-7 of 12.00 (SD=529). This suggests that in terms of baseline outcome scores, the populations seem similar. Summary statistics for a selection of linguistic features are also presented. These provide the mean and range information for scores of linguistic features that are relevant to the models in this chapter. As was the case in previous chapters, the scores refer to the percentage of language used by an individual in a therapy session that qualifies within a given language category or feature. As compared to the range of scores in the development data set, this was narrower in the validation set for six of the ten linguistic features presented. These were patient social language (LIWC), patient use of negations (LIWC), patient use of first person singular pronouns (LIWC), patient use of joviality language (PANAS-X based I2E query), therapist negative language (LIWC), and therapist positive language (PANAS-X based I2E query). For the last three of these the mean values were lower in the validation data set. Only mean patient use of first person singular pronouns was higher in the validation data set as compared to the development data set. All other mean and range measures were similar between the two datasets. These summary statistics suggest some differences in the language used between the two data sets.

**Table 9-3 Summary statistics of linguistic features present in validation models**

| Linguistic feature | Mean percentage score | SD | Min | Max |
|---|---|---|---|---|
| **Patient Social (LIWC)** | 0.94 | 0.78 | 0 | 5.56 |
| **Patient use of negations (LIWC)** | 2.05 | 0.91 | 0 | 6.99 |
| **Patient first person singular pronouns (LIWC)** | 7.34 | 1.91 | 2.21 | 16.42 |
| **Patient positive language (LIWC-based I2E query)** | 2.70 | 2.92 | 0 | 60.6 |
| **Patient joviality (Expanded PANAS-X category)** | 0.25 | 0.28 | 0 | 2.35 |
| **Therapist negative language (LIWC)** | 1.31 | 0.76 | 0 | 5.35 |
| **Therapist use of negations (LIWC)** | 0.76 | 0.47 | 0 | 3.32 |
| **Therapist Insight (LIWC)** | 3.15 | 1.10 | 0 | 7.25 |
| **Therapist positive language (Expanded PANAS-X based I2E query)** | 0.69 | 0.45 | 0 | 2.96 |
| **Therapist negative language (LIWC-based I2E query)** | 1.26 | 0.72 | 0 | 4.79 |

## 9.2 Validation results

I present here the results of using the models developed and presented in the previous chapter to predict each outcome score and compare these with the observed values. The performance of the model will be assessed using an estimated R-squared measure and calibration slope (Steyerberg et al., 2010) will also be presented for each model. The data set used for this validation was independent from that used for model development. It was collected at a later date (one year later) and involved patients from a different geographical location. The data set consisted of transcripts and case information for 376 patients who attended a total of 1667 appointments. Of the 376 patients in the data set 185 completed treatment, 171 dropped out of treatment and 20 were found to be unsuitable for the service or referred for treatment elsewhere.

### 9.2.1 Outcome 1 – PHQ-9 score before session

This model considered demographic, baseline and linguistic features associated with a therapy session as predictors for the PHQ-9 score recorded before the session. Table 9-4 below presents the summary statistics for both the predicted and observed values. Means were weighted by the number of recorded outcome scores.

**Table 9-4 Summary statistics of observed and predicted PHQ-9 scores before session**

| ***PHQ-9 score values*** | ***Weighted mean (by number of sessions attended)*** | ***Standard Deviation*** | ***Min*** | ***Max*** |
|---|---|---|---|---|
| ***Predicted*** | 9.58 | 6.68 | -3.17 | 22.76 |
| ***Observed*** | 9.76 | 6.53 | 0 | 27 |

An R-squared value was calculated as the proportion of the total variation in the outcome in the validation set explained by the models developed in Chapter 8. This was calculated by dividing the residual variance from predicted values by the total variance of observed values and subtracting this result from one. A large R-squared means the model explains a high proportion of the variability in the dependent variables with high and low predicted values indicating widely differing prognoses. The results of this calculation estimated an R-squared of 0.42 in this model. Thus 42% of the total variation in the outcome is explained by the model. Therefore the ability of the model to discriminate between patients with high and low scores is moderate. Plotting the residuals suggested a normal distribution of these, albeit with a small deviation at the lower end of the scale (Figure 9-1).

However, the baseline model was estimated to account for almost 45% of the variation in the outcome scores in the validation data, suggesting that the inclusion of the linguistic features does not improve model performance in this data set.

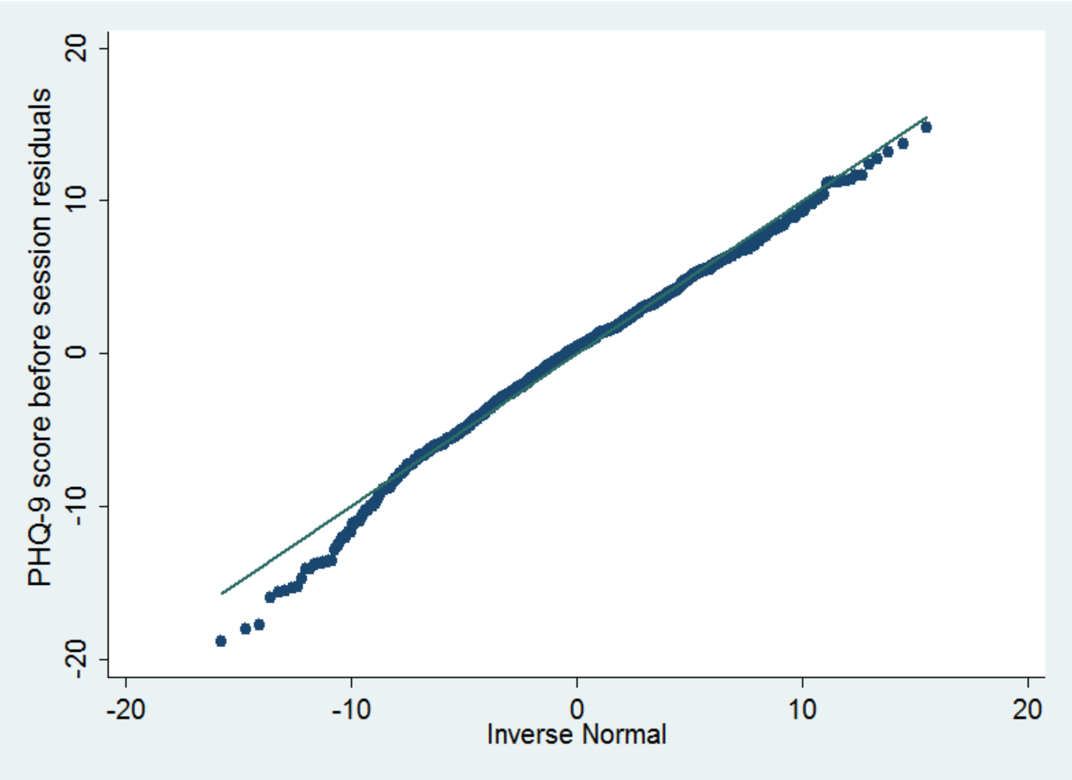


**Figure 9-1 Quantile normal plot of residuals from model predicting PHQ-9 score before a session**

A linear regression of the observed scores by the predicted values provides a calibration slope for this model of 0.93 (*b0* = 0.76), suggesting a reasonably well-calibrated model, reaching similar levels as during cross-validation in previous chapters. A calibration slope was also estimated with the inclusion of patient identity as a random effect in the model to account for the repeated measurements per individual in the validation data. In this version, the calibration slope was 0.98 (*b0*= 0.53), it supports the previous result, as both estimates are similar. Due to the repeated measurements, means by patient of the predicted and observed outcome values were calculated. These are presented in the scatter plot below to show the agreement between these values. Though the model comes across as quite predictive on average in this validation set, there is nonetheless some noise in the data, with predicted values spanning a majority of the observed range of scores. In a number of cases, the distance between the predicted and observed value suggests definite problems with the implementation of this prediction model in practice.

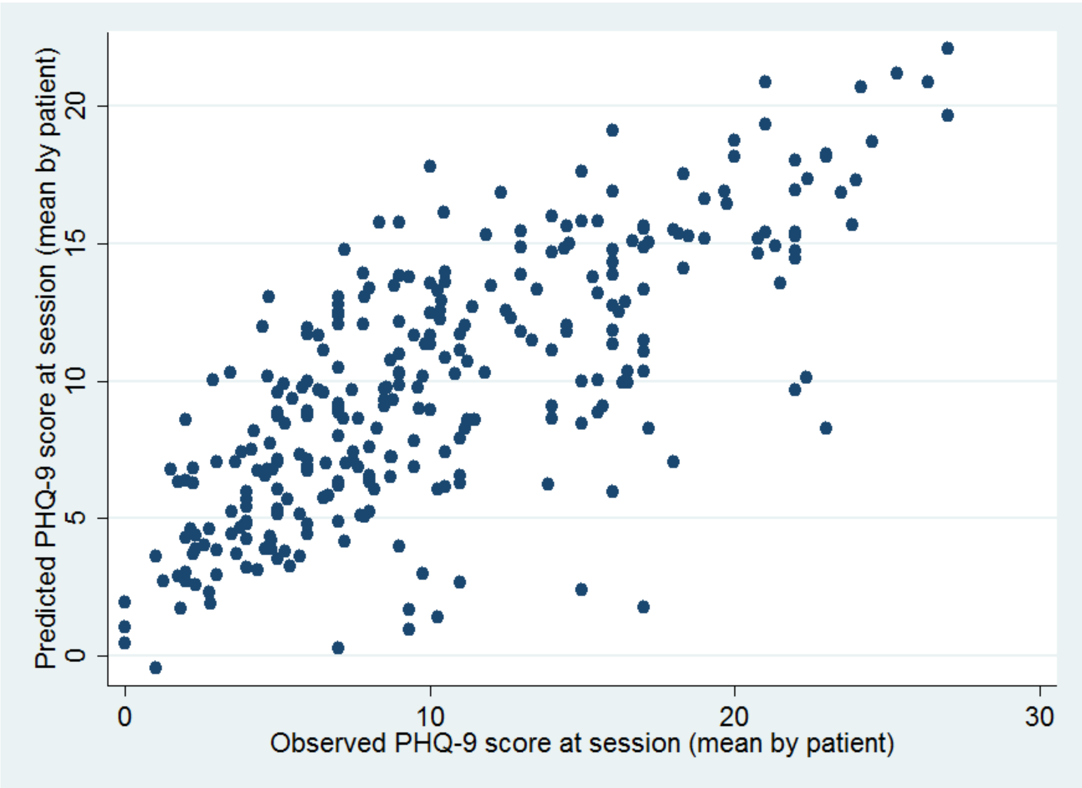


**Figure 9-2 Scatter plot of predicted and observed values of PHQ-9 score before session**

It is notable that there were some negative predicted values. In terms of the PHQ-9 scale, negative values are not possible. However, some PHQ-9 scores in the data are very low, with some individuals even reporting 0 as their PHQ-9 score. A model that underestimates this score may therefore predict a negative value. Manual checking and consideration of the predicted and observed scores suggested that cases where the outcome score was predicted to be negative are all cases in which the baseline PHQ-9 score is very low; 3 or below. As was demonstrated in previous chapters, the models presented are very reliant on baseline scores and a very low baseline score may lead to a negative outcome score. The presence of low baseline values presents further concerns with the dataset as patients were expected to present with a baseline PHQ-9 score of 10 or above when referred for treatment. This may be explained by the delay between referral for treatment and access to the service, which ranged between 2 and 147 days.

Following validation of the model above, the developed model was re-calibrated using the validation data set to explore how coefficient sizes may differ between the two data sets. The model was fitted on data from 1138 appointments involving 293 patients. Only variables that were statistically significant at the 15% level in the development set were put forward and only variables that were also statistically significant at the same level in the validation set were retained in the model. The model emerging from this process is presented below (see Table 9-5).

Four of the nine linguistic features that were included in the model fitted on the development data set also were when the model was tested on the validation data set. These were patient social language (LIWC), patient first person plural pronoun use, patient joviality (expanded PANAS-X), and therapist positivity (Expanded PANAS-X based I2E query). Of these, only patient social language was positively associated with outcome. However, the p-values attached to these associations were all around 0.70 or above, suggesting only moderate or weak evidence supporting the reality of these effects.

**Table 9-5 Results from re-calibrated model predicting PHQ-9 before session**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.61 | [ 0.54 ; 0.68] | <0.001 |
| ***Number of sessions*** | -0.71 | [ -0.81 ; -0.60 ] | <0.001 |
| ***Step group 2 – ref . group*** | | | |
| ***Step group 3*** | 2.36 | [ 1.19 ; 3.53 ] | <0.001 |
| ***Step group 3+*** | 5.91 | [ 3.96 ; 7.87 ] | |
| ***Patient Social language (LIWC)*** | 0.33 | [ -0.03 ; 0.69 ] | 0.069 |
| ***Patient First person plural pronouns*** | -1.12 | [ -2.54 ; 0.30 ] | 0.121 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.77 | [ -1.64 ; 0.06 ] | 0.069 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.51 | [ -1.05; -0.04 ] | 0.070 |
| ***Constant*** | 3.38 | [ 2.06 ; 4.70 ] | <0.001 |

These changes in the model after recalibration suggest some important differences between the linguistic features that were suggested to be associated with outcome in the development set and the validation set. Patient negative language (LIWC), patient use of negations, therapist certainty (LIWC), therapist positive language (LIWC-based I2E query) and agenda setting were all linguistic features that were not retained when the model was recalibrated. This suggests that though an association with outcome was suggested in the development set, this was not necessarily true in the validation set. However, the baseline and demographic features are maintained as strongly significant predictors of outcome in both datasets. The differences in the results attached to the linguistic features may point to weak predictors or differences in language use between the two populations.

When the same validation process was followed for the model developed with data from only patients who had completed their course of therapy, the performance of the model in the validation data was very similar. This model was fitted on data from 900 appointments involving 173 patients. The associated R-squared was 0.43, almost identical to that presented above and the calibration slope was 1.04 (*b0*=-0.24). The calibration slope was close to one but slightly larger.

The scatter plot of predicted and observed data points is not presented as this looks much the same as that presented above, with a slightly wider range of error between the predicted and observed PHQ-9 values. The recalibrated model is presented below (Table 9-6) as there were some differences there. Therapist positive language (Expanded PANAS-X based I2E query) was included in the model as it had been when the model was recalibrated with the full data set but with a high associated p-value. However, patient positive language (LIWC-based I2E query) was also included in this model, whereas this was not the case in the previous model.

**Table 9-6 Results from re-calibrated model predicting PHQ-9 before session - completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.64 | [ 0.54 ; 0.74] | <0.001 |
| ***Number of sessions*** | -0.71 | [ -0.82 ; -0.60 ] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.22 | [ -0.12 ; 2.56 ] | <0.001 |
| ***Step group 3+*** | 5.47 | [ 3.26 ; 7.68 ] | |
| ***Patient Positive language (LIWC-based I2E query)*** | -0.17 | [ -033 ; -0.01 ] | 0.040 |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.49 | [ -1.07; -0.10 ] | 0.103 |
| ***Constant*** | 3.38 | [ 2.06 ; 4.70 ] | <0.001 |

### 9.2.2 Outcome 2 – GAD-7 score before session

This model considered demographic, baseline and linguistic features associated with a therapy session as predictors for the GAD-7 recorded before the session.

Table 9-7 below presents the summary statistics for both the predicted and observed values. The means presented were weighted by the number of recorded outcome scores per individual. Though the mean values are very similar in the predicted and observed GAD-7 values, a larger spread of scores is suggested by the higher standard deviation in the observed values.

**Table 9-7 Summary statistics of predicted and observed GAD-7 score before session**

| *GAD-7 score values* | *Weighted mean (by number of sessions attended)* | *Standard Deviation* | *Min* | *Max* |
|---|---|---|---|---|
| ***Predicted*** | 9.41 | 3.88 | -1.56 | 18.94 |
| ***Observed*** | 9.29 | 4.96 | 0 | 21 |

As in the previous model, an R-squared value and calibration slope were calculated. The R-squared attached to this model was 0.31 suggesting that the developed model explained 31% of the variation in the outcome scores in the validation set. This is only slightly lower than the mean R-squared estimated through internal cross-validation in the previous chapter suggesting some loss of model fit but not a large amount. However, the baseline model fitted on the development data was also estimated to account for 31% of the variation in the outcome scores in the validation data set, suggesting no gain from the inclusion of the linguistic features in the model.

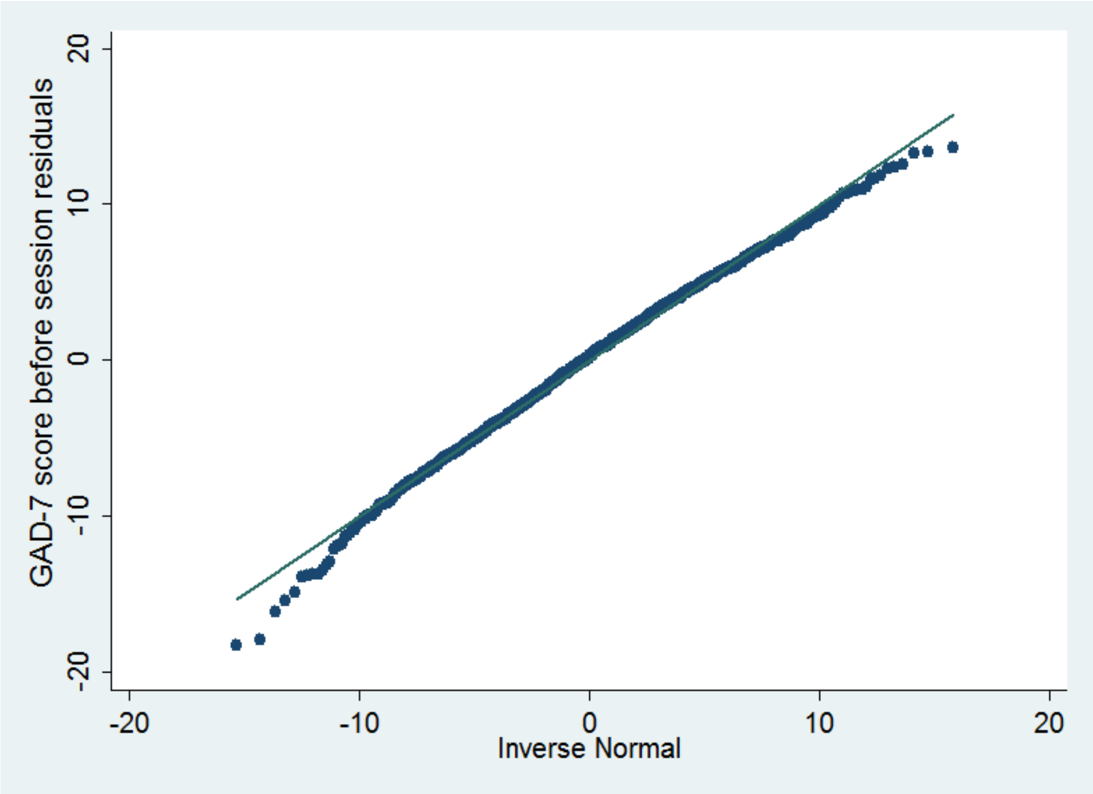


**Figure 9-3 Quantile normal plot of residuals from model predicting PHQ-9 score before a session**

The distribution of residuals was similar to the previous model; a normal distribution with a slight deviation at the extremes of the graph.

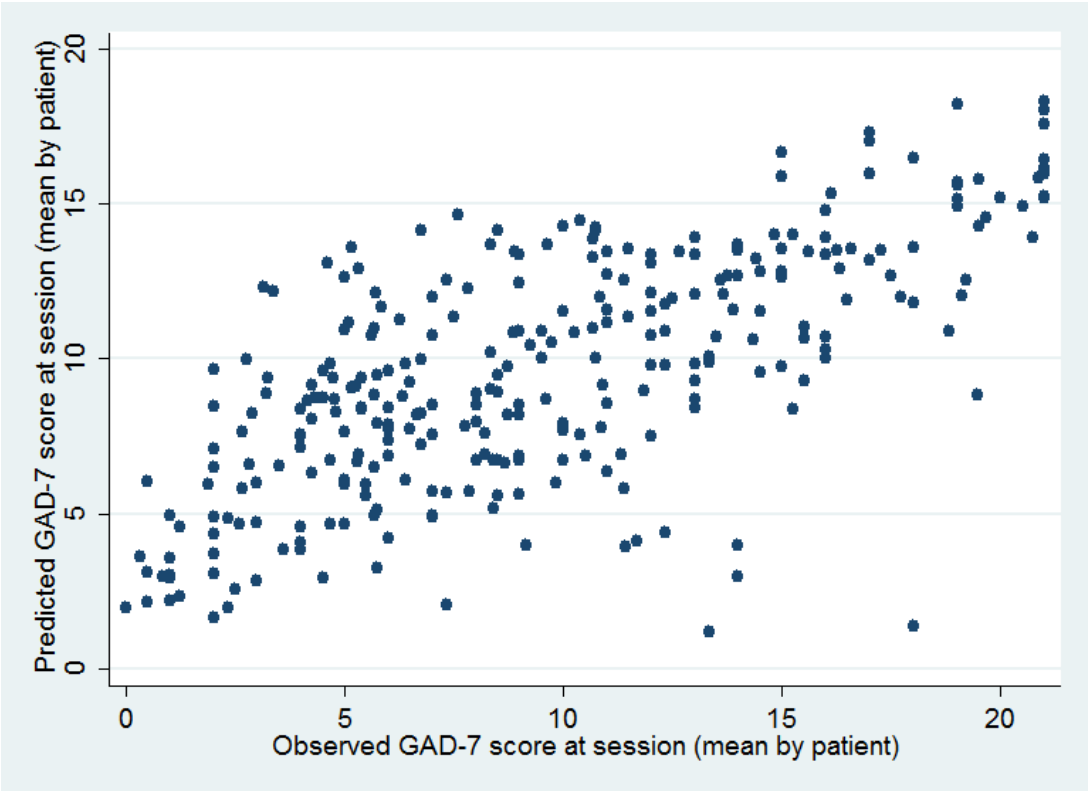


**Figure 9-4 Scatter plot of predicted and observed values of GAD-7 score before session**

The calibration slope associated with this model was 0.90 (*b0*=0.73), a lower score than was associated with the previous model and suggesting a weaker model. When this was estimated using a random effects model, however, it was 1.08 (*b0*= -0.84). This suggests a slightly better calibrated model.

A scatter plot of the mean predicted and observed outcome scores by individual also suggests a slightly weaker model than that presented above as a much wider range of predicted scores for each actual outcome score can be observed. This suggests ranges of error that would be of concern in any clinical implementation of this type of model.

**Table 9-8 Results from re-calibrated model predicting GAD-7 before session**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.51 | [ 0.44 ; 0.59] | <0.001 |
| ***Number of sessions*** | -1.50 | [ -1.83 ; -1.15 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.07 | [ 0.04 ; 0.10] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 2.89 | [ 1.75 ; 4.02] | <0.001 |
| ***Step group 3+*** | 6.28 | [ 4.37 ; 8.18 ] | |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.43 | [ -0.94 ; -0.08 ] | 0.102 |
| ***Therapist negative language (LIWC-based I2E query)*** | 0.47 | [ 0.14 ; 0.79 ] | 0.005 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.82 | [ -1.58 ; -0.03 ] | 0.034 |
| ***Constant*** | 4.99 | [ 3.45 ; 6.53 ] | <0.001 |

This model was re-calibrated on data from 1130 appointments attended by 293 patients. Re-calibration of this model with the validation set data suggests that three of the nine predictors were retained within the model. These were therapist negative language (LIWC-based I2E query), therapist positive language (Expanded PANAS-X based I2E query) and patient Joviality (Expanded PANAS-X category). All three features were associated with outcome in the same direction as they had been in the development set with similar but slightly larger associated coefficients. As was the case in the previous model, the baseline features were maintained as strong predictors of outcome in this model with the exception of diagnostic group that was no longer significantly associated with the GAD-7 outcome score. These results may indicate not only differences in language use between the populations but also in the attribution and significance of their provisional diagnoses.

When the same process was followed to validate and re-calibrate the model developed with only data from individuals who completed their course of therapy, the validation results were quite similar. This model was fitted on data from 896 appointments involving 172 patients. The R-squared associated with the complete cases model was 0.29, once again slightly smaller than that presented above. The associated calibration slope was 0.92 (*b0*= 0.15). Graphical observation of the residuals showed a very similar pattern as previously with slightly less deviation from the normal values at the extreme values in the complete cases dataset.

Results from the recalibration of the model (Table 9-9 below) suggest that two of the same linguistic features were retained. These were patient joviality (Expanded PANAS-X category) and therapist positive language (Expanded PANAS-X based I2E query). Both were associated with outcome with almost identical coefficients as those presented in Table 9-8 but neither association had a very low attaché p-value, again suggesting only moderate to weak evidence supporting them. Therapist insight (LIWC category) was statistically significantly associated with outcome in this model whereas it was not in the model including all patient cases. Therapist insight language was positively associated with GAD-7 score suggesting that higher levels of therapist insight language was associated with a higher GAD-7 score before a therapy session. Though this may seem counterintuitive, as insight is often associated with improvement in therapy, this higher insight levels may be a consequence of higher anxiety levels that require a therapist to use more skills, notably insight, in order to assist the patient. Further ideas around interpretation of these results in practical context will be put forward in later chapters.

**Table 9-9 Results from re-calibrated model predicting GAD-7 before session - completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.49 | [ 0.40 ; 0.59] | <0.001 |
| ***Number of sessions*** | -1.63 | [ -2.01 ; -1.27 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.08 | [ 0.05 ; 0.11] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.98 | [ 0.68 ; 3.29] | <0.001 |
| ***Step group 3+*** | 6.00 | [ 3.85 ; 8.15 ] | |
| ***Therapist positive language (Expanded PANAS-X based I2E query)*** | -0.44 | [ -1.00 ; 0.12 ] | 0.123 |
| ***Therapist Insight (LIWC)*** | 0.28 | [ 0.03 ; 0.52 ] | 0.028 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.80 | [ -1.63 ; 0.02 ] | 0.057 |
| ***Constant*** | 5.35 | [ 3.45 ; 6.53 ] | <0.001 |

### 9.2.3 Outcome 3 – PHQ-9 score before the next session

This model considered demographic, baseline and linguistic features associated with a therapy session as predictors for the PHQ-9 score recorded before the next session.

Table 9-10 below presents mean and range statistics for both the predicted and observed values. These suggest some difference in the distribution of observed and predicted values. The mean of observed values was lower than that of the predicted values but the range and spread of scores was wider.

**Table 9-10 Summary statistics for predicted and observed PHQ-9 score before next session**

| *PHQ-9 score values* | *Weighted mean (by number of sessions attended)* | *Standard Deviation* | *Min* | *Max* |
|---|---|---|---|---|
| ***Predicted*** | 10.32 | 4.67 | 0.002 | 24.35 |
| ***Observed*** | 8.95 | 6.76 | 0 | 27 |

The R-squared associated with the validation of this model was 0.375, suggesting that the developed model explained 37.5% of the variation in the outcome scores in the validation data. This is a weaker model than that predicting PHQ-9 score reported just before the therapy session but is nonetheless a useful model. The baseline model fitted on the development data set was estimated to account for almost 37% of the variation in the outcome scores in the validation data, suggesting limited gain from the inclusion of the linguistic features.

The calibration slope associated with this model was 0.90 (*b0* = -0.39), also suggesting a weaker model than that associated with outcome 1 – PHQ-9 score before a session. When fitted with a random effects model, the estimated calibration slope was 0.97 (*b0* = -1.12), suggesting a better calibrated model when random effects were taken into account. Observation of the distribution of the residuals in this model suggested this is very similar to that presented in the previous models; they follow the expected normal distribution for the most part with some deviation at the extremes.

A scatter plot of the mean predicted and observed scores by individual suggests a similar distribution of data points as was found in the validation data for the model predicting PHQ-9 score recorded before a session. There was however, more error in the predictions in this model, with a particularly large range of error when PHQ-9 score before the next session was equal to zero. These results are in line with the lower R-squared attached to this model and lower calibration slope.

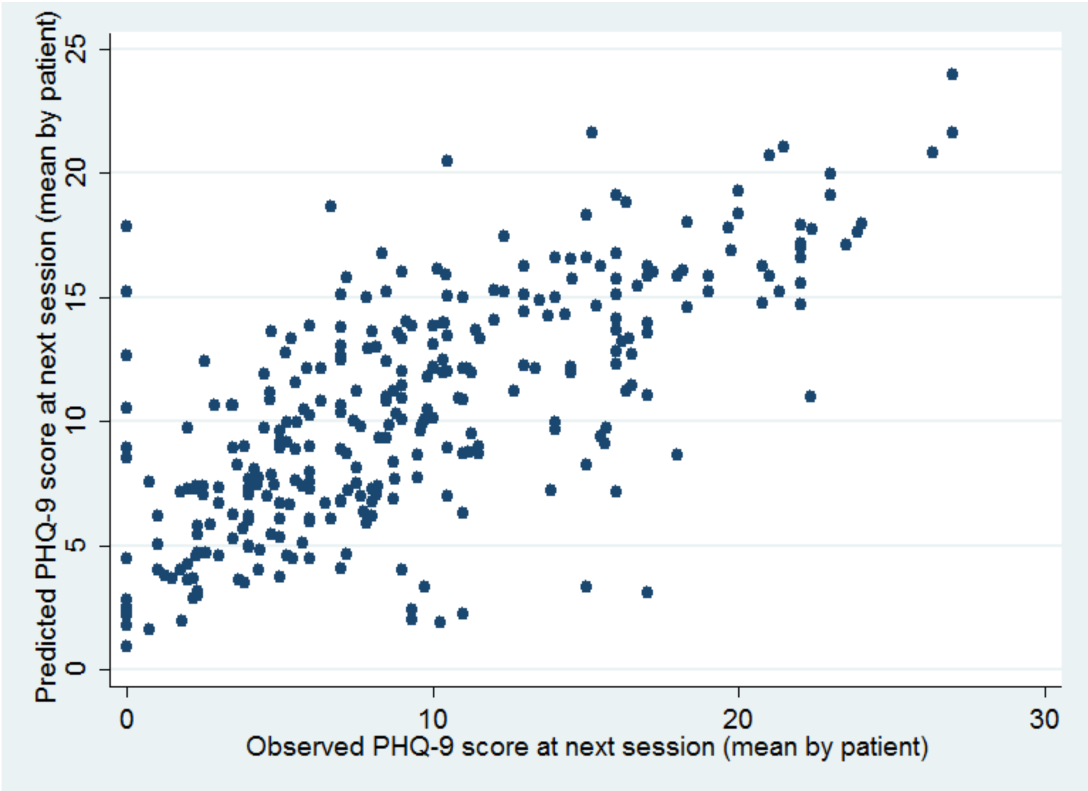


**Figure 9-5 Scatter plot of predicted and observed values of PHQ-9 score before session**

The model was refitted on the data from 908 appointments involving 204 patients. Re-calibration of the model on the validation dataset suggested that in this dataset, only two of the suggested linguistic features were retained in the model. These were therapist use of negations and patient expressions of joviality (expanded PANAS-X category). Both coefficients were approximately double the size they were in the originally developed model but maintained the same direction of association as previously. Therapist negations were positively and significantly associated with outcome suggesting that a higher level of therapist negation use was associated with a higher PHQ-9 score before the next session. A higher level of patient joviality (expanded PANAS-X category) in a session was associated with a lower PHQ-9 score reported before the next session, but there was less statistical evidence supporting this association than the previous one. The baseline and demographic predictors that were previously significantly associated with outcome were maintained as significant predictors in this model. The results of the full model can be found in Table 9-11.

**Table 9-11 Results from re-calibrated model predicting PHQ-9 score before next session**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.57 | [ 0.49 ; 0.65 ] | <0.001 |
| ***Number of sessions*** | -1.13 | [ -1.49 ; -0.78 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.05 | { 0.01 ; 0.08 ] | 0.016 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 2.07 | [ 0.85 ; 3.29 ] | <0.001 |
| ***Step group 3+*** | 5.02 | [ 2.95 ; 7.10 ] | |
| ***Therapist Negations (LIWC)*** | 0.87 | [ 0.27 ; 1.47 ] | 0.005 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.85 | [ -1.85 ; 0.13 ] | 0.090 |
| ***Constant*** | 2.46 | [ 1.05 ; 3.86] | 0.001 |

When the same process was followed to validate the model developed on only data from patients who completed their course of treatment, results were very similar. The associated R-squared was 0.38 and the calibration slope was 1.01 (*b0*=-0.60), this suggests a stronger model than that considering all cases, suggesting perhaps that individuals completing their course of treatment were more similar across dataset than those who didn't. The calibration slope estimated using a random effects model was 1.09 (*b0*=-1.46). This is consistent with the previous estimate. The model was then re-fitted on data from 769 appointments involving 172 patients. Re-calibration of the model with the complete cases in the validation dataset put forward the same predictor variables as in Table 9-11 with some small changes in coefficient values. Both predictors also had stronger statistical support with p-values below the lower 0.05 threshold. Two variables that had previously been associated with outcome in the development set were retained in the re-calibrated model. These were patient use of first person pronouns and therapist positive language (Expanded PANAS-X based I2E query).

**Table 9-12 Results from re-calibrated model predicting PHQ-9 score before next session – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.58 | [ 0.49 ; 0.68 ] | <0.001 |
| ***Number of sessions*** | -1.22 | [ -1.59 ; -0.85 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.05 | [ 0.02 ; 0.09 ] | 0.006 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.29 | [ -0.06 ; 2.65 ] | <0.001 |
| ***Step group 3+*** | 4.71 | [ 2.49 ; 6.93 ] | |
| ***Therapist Negations (LIWC)*** | 0.80 | [ 0.16 ; 1.44 ] | 0.014 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -1.15 | [ -2.21 ; -0.08 ] | 0.035 |
| ***Constant*** | 2.46 | [ 1.05 ; 3.86] | 0.001 |

These two sets of results suggest that though a number of linguistic features were significantly associated with outcome in the development set, this was not necessarily the case with a new data set. The variables that were statistically significant in both sets, however, may provide some interesting information about which elements of language may indicate treatment success. Patient Joviality, a category of words expressing happiness and enthusiasm, for example, may provide some indication towards a patient's feelings about their course of treatment.

### 9.2.4 Outcome 4 – GAD-7 score before the next session

This model considered demographic, baseline and linguistic features associated with a therapy session as predictors for the GAD-7 score recorded before the next therapy session.

Table 9-13 presents the descriptive statistics for both the predicted and observed values. As was the case in previous models, there appear to be some small differences between mean and range statistics for observed and predicted values. In this case the mean of observed values was lower than the mean of predicted values but there was a broader spread of values in the observed scores.

**Table 9-13 Summary statistics of predicted and observed GAD-7 score before next session**

| *GAD-7 score values* | *Weighted mean (by number of sessions attended)* | *Standard Deviation* | *Min* | *Max* |
|---|---|---|---|---|
| ***Predicted*** | 9.66 | 3.82 | -0.58 | 19.92 |
| ***Observed*** | 8.56 | 6.25 | 0 | 21 |

The R-squared associated with this model was 0.27, suggesting that the developed model explained 27% of the variation in the outcome score in the validation data. This is again slightly weaker than the model predicting the GAD-7 before the session. It was nonetheless a model with clear predictive value. However, the baseline model was estimated to account for 26% of the variation in the outcome scores in the validation data set suggesting only marginal improvement when linguistic features were included. The distribution of residuals was much the same as in the previously presented results, with perhaps less deviation from the expected values at the extremes in this case and therefore a more normal distribution of residuals. The calibration slope associated with this model was 0.87 (*b0* = -0.03), demonstrating once again a drop in the model's predictive accuracy as compared to the previous GAD-7 model and suggesting model over fitting. However, when this was estimated using a random effects model, the calibration slope was 0.99 (*b0* = -1.28), suggesting a very well calibrated model.

A scatter plot of the mean predicted and observed values by individual demonstrated visually the range of error found within this model. The results were clearly correlated but the range of predicted values associated with the observed values is wide. This would be cause for concern in any clinical implementation of this type of model.

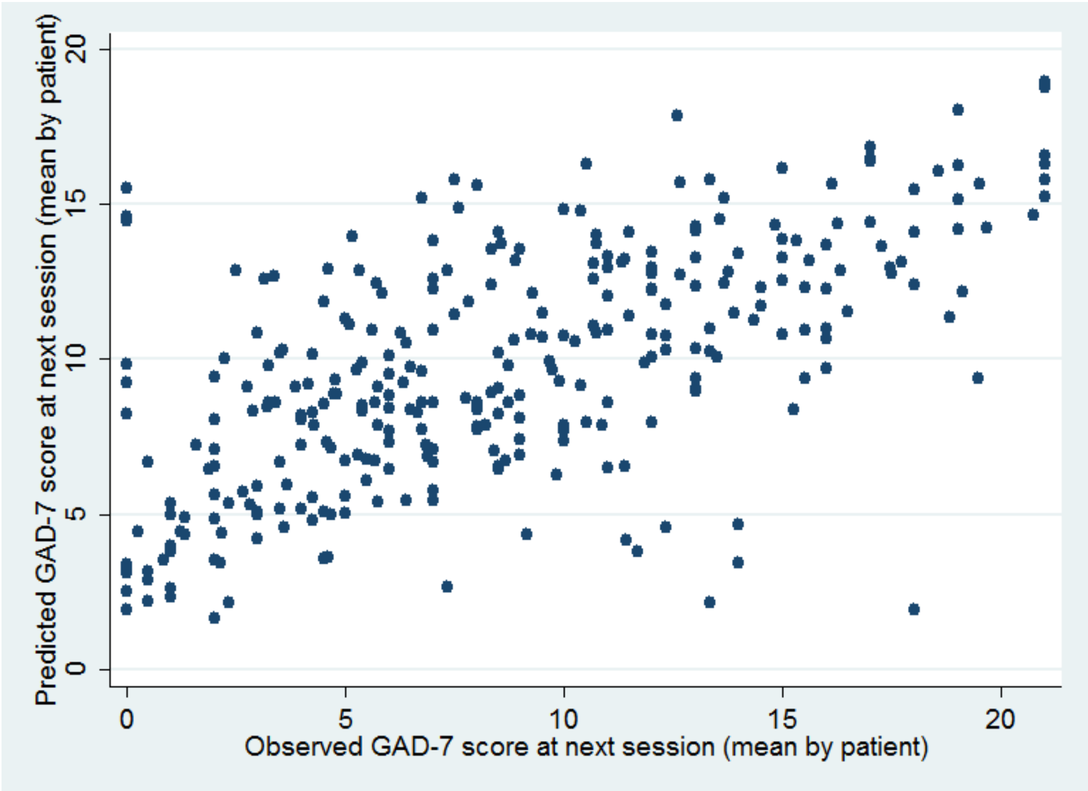


**Figure 9-6 Scatter plot of predicted and observed values of PHQ-9 score before next session**

As with the previous models, this model was re-calibrated and fit within the validation set (see Table 9-14) on data from 908 appointments involving 240 patients. Of the linguistic features that were put forward as candidate predictors, only two were retained within this model. These were therapist negative language (LIWC) and patient joviality (expanded PANAS-X). For both features, the direction of association was maintained. The coefficient associated with therapist positive language remained almost identical with *b* = 0.26 (95% CI = [ -0.09 ; 0.61 ] ,*p* = 0.147) but the coefficient associated with patient joviality was larger in this model *b* = -0.75 (95% CI = [ -1.70 ; 0.19 ], *p* = 0.118) as compared to *b* = -0.28 (95% CI = [ -0.56 ; -0.004 ], *p* = 0.046). However, the p-values suggest there is only weak evidence supporting these associations in this dataset. Additionally to the changes in linguistic features, the diagnostic group was not significantly associated with outcome in this model. This is in line with the previously presented model concerning GAD-7 score reported before a therapy session.

**Table 9-14 Results from re-calibrated model predicting GAD-7 score before next session**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.46 | [ 0.38 ; 0.55] | <0.001 |
| ***Number of sessions*** | -1.11 | [ -1.43 ; -0.77 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.04 | [ 0.004 ; 0.07] | 0.027 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 2.63 | [ 1.42 ; 3.84 ] | <0.001 |
| ***Step group 3+*** | 5.89 | [ 3.85 ; 7.93 ] | |
| ***Therapist Negative language (LIWC)*** | 0.26 | [ -0.09 ; 0.61 ] | 0.147 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.75 | [ -1.70 ; 0.19 ] | 0.118 |
| ***Constant*** | 3.06 | [ 1.55 ; 4.57 ] | <0.001 |

When the same validation process was followed using only data from patients who completed treatment, the results were similar. The associated R-squared was 0.29 but the calibration slope improved to reach 0.95 (*b0* = -0.07). When estimated using a random effects model, it went up to 1.06 (*b0* = -1.9), following a similar pattern as was seen in previous models. Table 9-15 presents the re-calibrated model fitting on a data set from 769 appointments involving 172 patients. This model included only the linguistic features that were retained in the model in both the development set and the validation set. These were the same features as in the full data set: therapist negative language (LIWC) and patient joviality (Expanded PANAS-X category). Both of these features are dictionary-based measures. As was the case in the previous model, patient joviality (Expanded PANAS-X category) was negatively associated with outcome suggesting that higher levels of joviality in patient language were associated with a lower GAD-7 score reported before the next session. An opposite association was suggested for therapist negative language (LIWC) with higher levels of therapist negative language being associated with a higher GAD-7 score before the next session. However, as was also the case above, high p-values bring into question the reality of this association.

**Table 9-15 Results from re-calibrated model predicting GAD-7 score before next session – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.48 | [ 0.38 ; 0.58] | <0.001 |
| ***Number of sessions*** | -1.28 | [ -1.63 ; -0.93 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.06 | [ 0.02 ; 0.09] | 0.002 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.73 | [ 0.42 ; 3.04 ] | <0.001 |
| ***Step group 3+*** | 5.39 | [ 3.25 ; 7.53 ] | |
| ***Therapist Negative language (LIWC)*** | 0.29 | [ -0.09 ; 0.66 ] | 0.137 |
| ***Patient Joviality (Expanded PANAS-X category)*** | -0.97 | [ -1.98 ; 0.03 ] | 0.058 |
| ***Constant*** | 3.69 | [ 1.99 ; 5.38 ] | <0.001 |

### 9.2.5 Outcome 5 – PHQ-9 score at end of treatment

The same process for external validation was followed for the models developed to predict final therapy outcome based on demographics, baseline scores and linguistic features during the first two treatment sessions.

Table 9-16 presents the mean and range statistics of the predicted and actual end of treatment PHQ-9 scores.

**Table 9-16 Summary statistics of predicted and observed final PHQ-9 score**

| ***Final PHQ-9 score*** | ***Mean*** | ***Standard Deviation*** | ***Min*** | ***Max*** |
|---|---|---|---|---|
| ***Predicted*** | 7.10 | 3.25 | -1.63 | 16.15 |
| ***Observed*** | 6.10 | 6.49 | 0 | 27 |

The estimated R-squared summarizing explained variation over total variation in outcome scores associated with this model was 0.284. This suggests that the developed model explained 28.4% of the variation in the end of treatment PHQ-9 scores in the validation data. However, the baseline model was estimated to account for almost 29% of the variation in the end of treatment PHQ-9 scores, suggesting no gain from the inclusion of the linguistic features in this model. The calibration slope emerging from the regression of predicted and observed values was 1.07 (*b0*= -1.54). The summary values of predicted and observed end of treatment PHQ-9 scores suggest that the range of predicted values is narrower than that of observed values with the maximum value over 10 points lower despite a higher mean score. The residual values were broadly normally distributed but with some deviation, primarily at the lower end of the residuals.

A scatter plot of the mean predicted and observed values by individual suggests that, though the scores are clearly correlated, there is a large amount of noise in the plot. As in previous models, the larger error seems to congregate around the lower observed values, in particular where these were zero. The error in predicted scores suggested by this plot suggests some significant problems with the model if it were to be considered for application in practice.

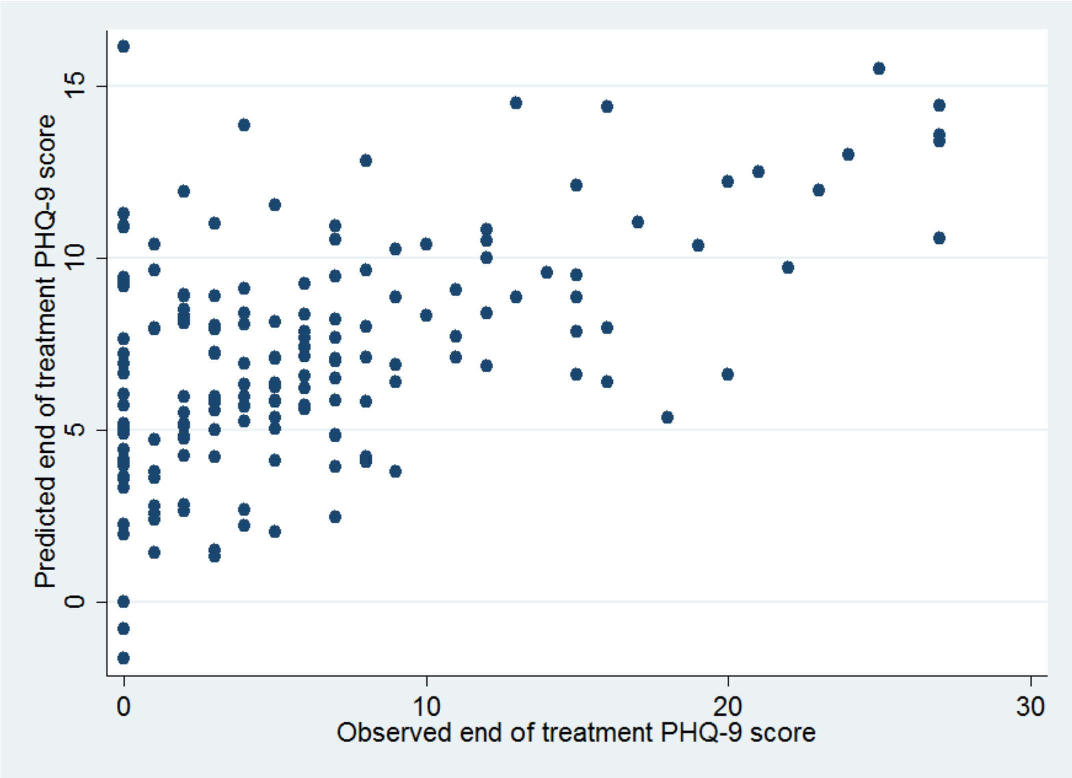


**Figure 9-7 Predicted and observed end of treatment PHQ-9 scores**

This model was also re-calibrated within the validation data set. The results of the re-calibration suggest that only two of the four linguistic features that were previously retained in the model, also were in the validation dataset. These were patient positive language (expanded PANAS-X-based I2E query) and patient negation use. In both cases these refer to the mean scores for these features during the first two treatment sessions. The coefficients associated with each of the predictors increased as compared to the model fitted on the development data but the direction of association was maintained and both appeared to be significantly associated with outcome. Patient social language (LIWC) and therapist positive language (LIWC-based I2E query) were not retained in these models whereas they had been previously. As with previous models, these results suggest both similarities and differences in the associations between linguistic features and outcome scores. Though a number of the linguistic features are not statistically significant within this data set, others have been maintained, supporting the strength of the association between these and outcome.

**Table 9-17 Regression results predicting final PHQ-9 score**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.57 | [ 0.44 ; 0.71 ] | <0.001 |
| ***Patient Positive language (Expanded PANAS-X based I2E query)*** | -2.78 | [ -5.31 ; -0.25 ] | 0.031 |
| ***Patient Negations (LIWC)*** | 1.33 | [ 0.19 ; 2.46 ] | 0.022 |
| ***Constant*** | -1.69 | [ -4.95 ; 1.57 ] | 0.307 |

### 9.2.6 Outcome 6 – GAD-7 score at end of treatment

This model considered demographic, baseline and linguistic features early in therapy as predictors for the final GAD-7 score reported at the end of treatment.

Table 9-18 below presents the mean and range statistics for both the predicted and observed values of the final GAD-7 score reported.

**Table 9-18 Summary statistics of predicted and observed final GAD-7 score**

| ***GAD-7 score values*** | ***Mean*** | ***Standard Deviation*** | ***Min*** | ***Max*** |
|---|---|---|---|---|
| ***Predicted*** | 9.96 | 2.65 | 0.02 | 14.31 |
| ***Observed*** | 5.54 | 5.47 | 0 | 21 |

The R-squared associated with the predicted values was 0.16. This suggests that the developed model explained 16% of the variation in the end of treatment GAD-7 scores in the validation set. This is a much lower value than that put forward in the model fitted on the development data. Although it is a useful result with a significant amount of variation explained, improvements would need to be made prior to this type of model being applied in practice. Furthermore, the baseline model was estimated to account for 17% of the variation in the end of treatment GAD-7 scores in the outcome data, suggesting a model that performed better without the inclusion of the linguistic features. The calibration slope associated with the model was in line with this at 0.83 (*b0*= -0.29), suggesting a less precise and weaker model than that presented above.

Graphical observation of the residuals against expected normal values suggested that there was some deviation from the normal distribution of residuals at the extremes of residual values. This was particularly the case at the lower end, suggesting possible underestimation of a number of final GAD-7 scores. These were nonetheless close to a normal distribution.

The scatter plot of observed and predicted values shows again that despite a correlation between the observed and predicted values, there is clearly a large amount of error in the predictions from the model, a cause for concern in any potential clinical implementation of this model.

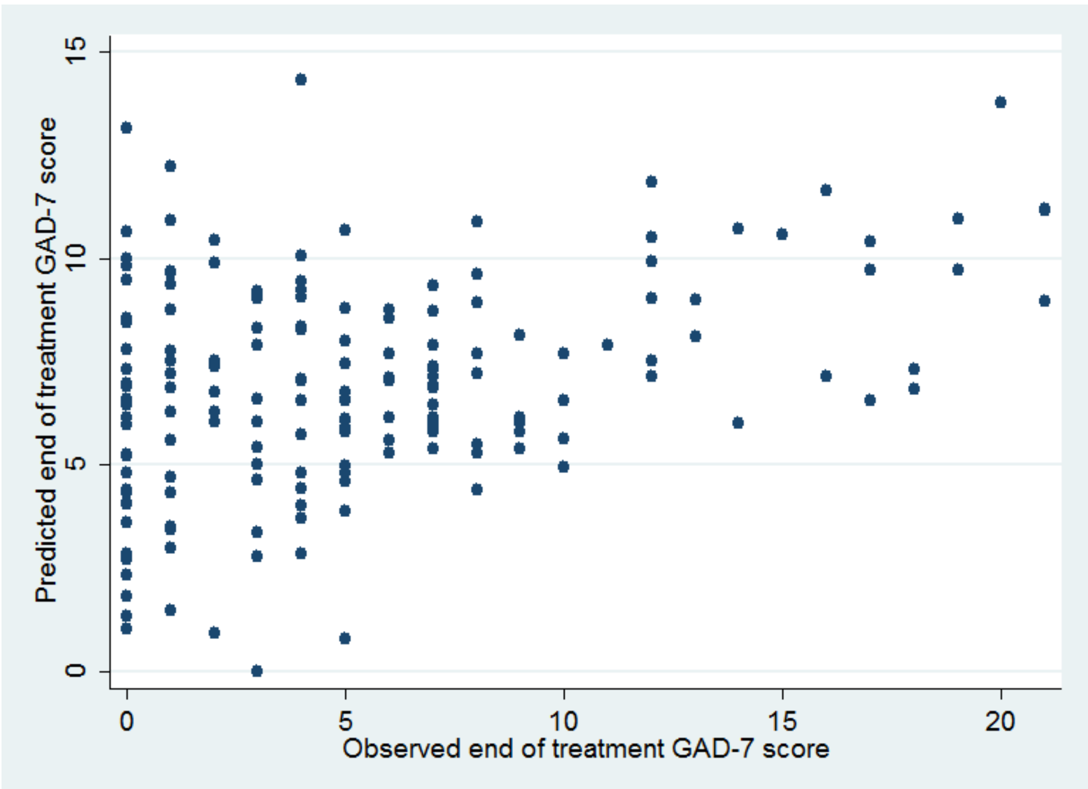


**Figure 9-8 Scatter plot of predicted and observed end of treatment GAD-7 scores**

This model was also re-calibrated with the validation dataset. The results of this re-calibration can be found below. The results suggest that among the linguistic features tested only patient negation use early in treatment was retained in the model of GAD-7 score at the end of treatment and with a p-value that suggest only modest evidence of the effect. These results suggest that only one of the associations between mean linguistic feature scores early in treatment and final GAD-7 score that had been statistically significant was maintained when tested on an external dataset and raises some important points for discussion in later chapters. The maintenance of only two predictors from the developed model may also explain the weaker model evidenced by the R-squared, calibration slope and range of error.

**Table 9-19 Regression results predicting final GAD-7 score**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.39 | [ 0.26 ; 0.52 ] | <0.001 |
| ***Patient Negations (LIWC)*** | 1.00 | [ -0.03 ; 2.03 ] | 0.058 |
| ***Constant*** | -1.10 | [ -3.71 ; 1.51] | 0.407 |

## 9.3 Overview of results

Throughout this chapter, there has been a recurring pattern of models showing poorer performance and calibration slopes when tested in an external data set (the validation set) as compared to those presented in the previous data set. Comparison of R-squared values from baseline models and models including linguistic features suggests very little improvement associated with the inclusion of linguistic features. Nonetheless, calibration was satisfactory in most cases. The pattern of stronger models associated with PHQ-9 outcomes scores and GAD-7 outcome scores reported just before a session (as opposed to models predicting the following outcome score) was maintained through validation. However, it seemed that, particularly in terms of model calibration, models developed and tested on the smaller data sets including only data from patient who had completed treatment were more stable. This could be associated with the smaller number of predictors or indicate that patients who completed treatment in both data sets were more similar than the two populations as a whole.

Finally, in terms of re-calibrated models, in each case some of the statistically significant predictors in the development set were not sso in the validation set, suggesting differences in language use between the two populations and the association with outcome. It is, however, important to note that a number of predictors were maintained in the validation set models. Measures of negative and positive language, particularly as used by patients, often persisted in the re-calibrated models. Patient Joviality (expanded PANAS-X category) was present in the majority of developed and re-calibrated models. This may indicate the importance of patient happiness and enthusiasm, and expression of this in treatment, as an indication of therapy progress and likelihood of success. This idea, and others put forward by the results described in this and previous chapters, will be further discussed in later chapters.

# Chapter 10. Clinical outcomes

This chapter will report on two sets of results with more direct clinical implications than those reported in previous chapters. The first considers whether the demographic, baseline and linguistic features reported on throughout this project are useful in the prediction of recovery from mental illness as defined within the IAPT framework. The second considers drop-out from treatment as a process outcome and how these same features may affect the likelihood of an individual dropping out of their course of treatment.

## 10.1 IAPT defined Recovery

Within IAPT, the concept of ‘caseness’ is used to define whether an individual would benefit from psychological therapy and whether they have or have not recovered from psychological disorder at the end of a course of treatment. An individual is considered in ‘caseness’ if they report a PHQ-9 score of 10 or above or a GAD-7 score of 8 or above. Of the 207 patients in the development data set who completed treatment and have recorded baseline scores, 35 were found to have PHQ-9 and GAD-7 scores below the ‘caseness’ threshold. In the validation data set, this number was 23 out of the 174 patients who completed treatment and have a recorded baseline score.

Individuals are also considered to have recovered at the end of treatment if they report a PHQ-9 score of under 10 and a GAD-7 score of under 8. This threshold is that applied when reporting official statistics and is therefore used as an indication of success of treatment. Table 10-1, below, provides overall recovery frequencies for all patients who completed treatment.

**Table 10-1 GAD-7 and PHQ-9 based recovery frequencies**

| *Recovery outcome* | | *Frequency* | *Percent.* | *Total frequency* |
|---|---|---|---|---|
| *PHQ-9 based recovery* | **Recovered** | 171 | 73.4 | 233 |
| | **Not recovered** | 62 | 26.6 | |
| *GAD-7 based recovery* | **Recovered** | 166 | 71.6 | 233 |
| | **Not-recovered** | 66 | 28.4 | |

The models that are presented below are logistic regression models that consider these binary definitions of recovery as an outcome (PHQ-9 recovery and GAD-7 recovery). For both PHQ-9 and GAD-7 recovery, a baseline model will first be presented followed by the combined model that included predictors in each set of linguistic features that were suggested to be associated with outcome.

## 10.2 PHQ-9 based recovery

### 10.2.1 Baseline model

This model considers the baseline PHQ-9 score, total number of sessions attended, age group, step group, diagnostic group and gender as potential predictors of binary recovery at the end of treatment.

**Table 10-2 Results of logistic regression prediction of PHQ-9 score based recovery from baseline features**

| *Predictors* | *Odds Ratio* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9 score*** | 0.84 | [ 0.80 ; 0.90 ] | <0.001 |
| ***Total sessions attended*** | 0.90 | [ 0.79 ; 1.02 ] | 0.108 |
| ***Constant*** | 53.33 | [13.96 ; 203.83] | <0.001 |

The results suggest that two baseline predictor variables were associated with outcome in this model. These were the baseline PHQ-9 score and the total number of sessions attended by the patient. Baseline PHQ-9 score was associated with outcome with an odds ratio of 0.85 (95% CI = [0.80 ; 0.90], $p < 0.001$). This suggests that for a one unit increase in PHQ-9 score, the odds of a patient recovering were 15% lower, on average. The total number of sessions attended was associated with outcome with an odds ratio of 0.90 (95% CI = [0.79 ; 1.02], $p = 0.108$) suggesting that for every session attended, the likelihood of recovery decreased by 10%, on average, however the high p value attached to this association suggests caution in interpretation.

The c-statistic estimated for this model was 0.78 (95% CI = [0.71 ; 0.84], suggesting a reasonably strong model when including only baseline PHQ-9 and number of appointments attended.

### 10.2.2 Combined model

This model considers the same binary recovery outcome, based on PHQ-9 score as above, but includes measures of language use in the first two treatment sessions as candidate predictor variables. The association between these mean linguistic features early in therapy and PHQ-9 based recovery was tested in stages, with each set of features being tested in turn. Only the results of the combined model will be presented here.

**Table 10-3 Results of logistic regression prediction of PHQ-9 score based recovery from baseline and linguistic features**

| *Predictors* | *Odds Ratio* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9 score*** | 0.84 | [ 0.79 ; 0.90 ] | <0.001 |
| ***Total sessions attended*** | 0.90 | [ 0.78 ; 1.02 ] | 0.091 |
| ***Therapist positive language (LIWC-based I2E query)*** | 2.37 | [ 1.29 ; 4.38 ] | 0.006 |
| ***Patient negations (LIWC)*** | 0.59 | [ 0.36 ; 0.99 ] | 0.045 |
| ***Patient insight (LIWC)*** | 1.52 | [ 1.02 ; 2.28 ] | 0.040 |
| ***Constant*** | 8.59 | [ 0.81 ; 91.59 ] | 0.075 |

The results of this combined model suggest that three linguistic features were statistically associated with outcome after all previous predictors were tested together. These were therapist positive language (LIWC-based I2E query), patient negation use (LIWC category) and patient insight (LIWC category). In the case of therapist positive language (LIWC-based I2E query), the odds ratio associated with this feature was 2.37 (95% CI [1.29 ; 4.38], $p$ = 0.006), putting forward evidence of a strong effect of therapist positive language early in treatment. This suggests that for a one per cent increase in mean therapist positive language use early in therapy, a patient's odds of recovery was 2.37 times higher, on average. Patient negation use had the opposite effect on odds of recovery with an odds ratio of 0.59 (95% CI = [0.36 ; 0.99], $p$ = 0.045), suggesting that patient odds of recovery decreased by 41% for a one per cent increase in mean use of negations in patient language early in therapy.

The estimated c-statistic associated with this model was 0.82 (95% CI = [0.76 ; 0.88]). This is an increase of 0.04 when compared to the baseline c-statistic (of 0.78).

### 10.2.3 Testing on validation data set

These models were tested on the data in the validation set following a similar process as in previous regression models. The parameters of the model estimated in the development data set were used to estimate the prognostic index associated with each case in the validation data set. The c-statistic was then estimated as an indication of model performance. When the combined model was tested on the data in the validation set, consisting of 159 patient cases, the associated c-statistic was 0.80 (95%CI = [0.72 ; 0.88]). This result alone suggests that the performance of the model was maintained in the new dataset. However, the baseline model appeared to perform better in the validation set, with the area under the ROC curve estimated as 0.83 (95% CI = [0.76 ; 0.90 ]).

A calibration slope of the developed model was also estimated by including the prognostic index of PHQ-9 based recovery as sole covariate in a logistic regression of recovery. This slope was estimated at 0.91 ($b0$= -0.04).

When the model was re-calibrated within the validation data set none of the linguistic features were significantly associated with outcome. Baseline PHQ-9 score and the total number of sessions attended by the patient were the only predictors significantly associated with outcome in the re-calibrated model. This may explain the lower performance of the complex model as it included predictors that may have reduced prediction accuracy. Together, these results suggest that the inclusion of the linguistic features in the model does not improve its performance when tested on an independent dataset.

The association between the total number of sessions and recovery is interesting and we may expect a higher number of treatment sessions to be associated with a higher likelihood of recovery. However, there is evidence to suggest that number of sessions attended is not closely associated with recovery in practice (Stiles, Barkham, & Wheeler, 2015). An explanation put forward by Stiles et al. (2015) for this is the idea of ‘responsive regulation of treatment duration’ by which patients and therapists agree to end treatment when the therapeutic gains are deemed good enough. It is possible that patients with more severe mental health issues tend to be offered and attend a higher number of total treatment sessions and that given the higher severity (and consequently PHQ-9 or GAD-7 score) at the beginning of treatment, these patients are less likely to have dropped below the threshold for 'caseness' and therefore recovery by the end of treatment. Considering a change score as an outcome in future analyses would help determine if this is the effect occurring here.

## 10.3 GAD-7 based recovery

### 10.3.1 Baseline model

This model considers the baseline GAD-7 score, total number of sessions attended, age group, step group, diagnostic group and gender as potential predictors of GAD-7 based recovery at the end of treatment.

**Table 10-4 Results of logistic regression prediction of GAD-7 score based recovery from baseline features**

| ***Predictors*** | ***Odds Ratio*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7 score*** | 0.83 | [ 0.77 ; 0.89 ] | <0.001 |
| ***Constant*** | 25.26 | [ 8.97 ; 71.17 ] | <0.001 |

The results of this model suggest that only the baseline GAD-7 score was significantly associated with recovery with an associated odds ratio of 0.83 (95% CI = [0.77 ; 0.89 ], $p$ < 0.001), suggesting that for every point on the baseline GAD-7 score, a patient was, on average, 17% less likely to recover.

The estimated c-statistic, associated with this model was 0.70 (95% CI = [0.62 ; 0.79]) suggesting that there is a 70% chance that a randomly chosen patient who recovered has a higher prediction of recovery according to the model than a patient who did not recover.

### 10.3.2 Combined model

This model considers the same binary recovery outcome, based on GAD-7 score as above, but includes measures of language use in the first two treatment sessions as candidate predictor variables. The association between these mean linguistic features early in therapy and GAD-7 based recovery was tested in stages, with each set of features being test in turn. Only the results of the combined model will be presented here.

**Table 10-5 Results of logistic regression prediction of GAD-7 score based recovery from baseline and linguistic features**

| *Predictors* | *Odds Ratio* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ-9 score*** | 0.82 | [ 0.76 ; 0.89 ] | <0.001 |
| ***Therapist positive language (LIWC-based I2E query)*** | 1.91 | [ 1.06 ; 3.44 ] | 0.032 |
| ***Patient negations (LIWC)*** | 0.69 | [ 0.43 ; 1.26 ] | 0.139 |
| ***Patient Joviality (Expanded PANAS-X category)*** | 3.84 | [ 1.68 ; 8.80 ] | 0.001 |
| ***Constant*** | 4.49 | [ 0.54 ; 37.15 ] | 0.164 |

The results of this combined model suggest that only three linguistic features were statistically significant at the 15% level after all features were tested. These were patient use of negations (LIWC category), therapist positive language (LIWC-based I2E query) and patient joviality (expanded PANAS-X category). Both therapist positive language and patient joviality were associated with recovery with odds ratios above 1, suggesting that higher mean levels of these linguistic features early in treatment were associated with a higher likelihood of recovery at the end of treatment. This was in line with many of the previously presented models that suggest that positive language in both the patient and therapist were associated with better outcomes (lower GAD-7 score) both during and at the end of treatment.

Patient use of negations was also associated with recovery, as was the case in the PHQ-9 based recovery model. The associated odds ratio was 0.69

(95% CI = [0.43 ; 1.26 ], *p* = 0.139) suggesting a 31% lower likelihood of recovery, on average, for a one percent increase in mean negation use in the first two treatment sessions, however this results should be interpreted cautiously given the high associated significance value. Alongside the language features, the baseline GAD-7 score was significantly associated with outcome with an almost identical odds ratio as was presented in the baseline model above.

The c-statistic associated with this model was estimated to be 0.81 (95% CI = [0.75 ; 0.87]), an improvement on that estimated in the baseline model.

### 10.3.3 Testing on validation data set

As was the case for PHQ-9 based recovery, these models were tested on the data in the validation set following a similar process as in previous regression models. When the combined model was tested on the data in the validation set, consisting of 158 patient cases, the associated c-statistic was 0.70 (95% CI = [0.61 ; 0.79]). However, when the baseline model was tested in the same way, the associated area under the curve was 0.70 (95% CI = [0.60 ; 0.80]). This suggests no gain from the inclusion of the linguistic features in this model. Additionally, the estimate of the calibration slope of the combined model was 0.60 (*b0* = 0.74), suggesting a poorly calibrated model when applied to this data set.

These results were further supported when the model was re-calibrated within the validation data set. As was the case with the PHQ-9 based measure of recovery, none of the linguistic features were significantly associated with outcome, with only baseline GAD-7 score being maintained as a statistically significant predictor in the model. Together, these results suggest that the model of GAD-7 based recovery fitted in the development set performed poorly when tested on the validation set and that the inclusion of linguistic features in this model did not improve it.

## 10.4 Drop-out from treatment

The second form of clinical outcome considered in this project was drop-out. There was some uncertainty over the definition of drop-out in this data set as the therapy provider (Ieso Digital Health) put forward a different definition of drop-out than IAPT. Within IAPT, a patient is considered to have completed treatment once they have attended a minimum of two treatment sessions after an assessment session. However, within Ieso, a patient is only considered to have completed therapy upon discharge and mutual agreement with their therapist. A patient who simply does not return for their next therapy appointment without being referred to another service is considered to have dropped out of treatment. As this is the more clinically useful definition of drop-out for the service, this was the definition applied in the following analysis.

Analysis of variables that may have an association with likelihood of drop-out was carried out using Cox proportional hazards survival analysis. This form of analysis considers the time to an event occurring or not, in this case, drop-out, and estimates the association between that event occurring and the variables included for analysis. The results will be expressed in terms of hazard ratio, meaning that a hazard ratio above one suggests that the event is more likely to occur and a hazard ratio below one suggests it is less likely to occur. A model considering baseline variables and a model considering linguistic features will be presented for both the development and validation set. The model developed in the validation set was not based on that in the development set as the aim of these models was not predictive but explanatory. The goal in this analysis was to explore the associations between linguistic features and outcome in order to determine whether language features used in a treatment session were associated with drop-out following that session. Ultimately, a long-term goal would be to influence likelihood of drop-out rather than accurately predict it, therefore explanatory analysis is more relevant (Shmueli, 2010).

### 10.4.1 Development set

Data from 473 individuals were included in this analysis, with 100 failures (drop-out events) within this dataset. The time at risk was 22,974 days.

#### 10.4.1.1 Baseline model

The analysis of drop-out patterns was initially carried out considering non-linguistic features; demographic information and recorded outcome scores for each session.

**Table 10-6 Results of Cox Proportional Hazards model for drop-out from treatment from baseline measures.**

| *Covariates* | *Hazard Ratio* | *95% CI* | *P* |
|---|---|---|---|
| ***PHQ-9 score*** | 1.07 | [ 1.03 ; 1.10 ] | <0.001 |
| ***Number of sessions attended*** | 0.59 | [ 0.47 ; 0.76 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.06 | [ 0.63 ; 1.78 ] | 0.016 |
| ***Step group 3+*** | 0.42 | [ 0.19 ; 0.91 ] | |

The results of this analysis suggest that three variables were statistically significant in their association with drop-out. These were the number of sessions attended (to date), the step group and the PHQ-9 score at a session. The hazard ratio associated with number of sessions attended was 0.59 (95% CI = [0.47 ; 0,76], $p$ < 0.001). This suggests that for every session a patient attended, their likelihood of dropping out of treatment was lowered by 59%, on average. The hazard ratio associated with PHQ-9 score was 1.08, suggesting that for every point higher on the reported PHQ-9 score, a patient was, on average, 8% more likely to drop-out after the session.

A global test of Schoenfeld residuals was not significant, suggesting that the proportional hazards assumption was respected in this model.

#### 10.4.1.2 Combined linguistic features model

This analysis was carried out on the same set of data with the inclusion of linguistic features. These were originally tested in sets and the combined results will be presented here.

**Table 10-7 Results of Cox Proportional Hazards model for drop-out from treatment from baseline measures and linguistic features**

| *Covariates* | *Hazard Ratio* | *95% CI* | *P* |
|---|---|---|---|
| ***PHQ-9 score*** | 1.05 | [ 1.02 ; 1.08 ] | 0.002 |
| ***Number of sessions attended*** | 0.51 | [ 0.39 ; 0.66 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.99 | [ 0.59 ; 1.67 ] | 0.020 |
| ***Step group 3+*** | 0.40 | [ 0.18 ; 0.87 ] | |
| ***Patient typing rate (words per minute)*** | 0.96 | [ 0.93 ; 0.99 ] | 0.020 |
| ***Patient certainty (LIWC)*** | 1.28 | [ 0.99 ; 1.66 ] | 0.058 |
| ***Patient negation use (LIWC)*** | 1.27 | [ 1.07 ; 1.50 ] | 0.005 |
| ***Patient Guilt (Expanded PANAS-X category)*** | 5.80 | [ 1.86 ; 18.08 ] | 0.002 |
| ***Therapist Positive language (Expanded PANAS-X-based I2E query)*** | 0.64 | [ 0.42 ; 0.98 ] | 0.041 |
| ***Therapist Negative language (Expanded PANAS-X-based I2E query)*** | 0.63 | [ 0.40 ; 1.00 ] | 0.054 |

The results suggest that, in addition to the variables in the baseline model, six linguistic features were statistically significantly associated with outcome in this analysis. Four of these were patient features and two were therapist features. Within the patient features were patient typing rate, patient use of negations (LIWC), patient certainty (LIWC) and patient guilt (expanded PANAS-X category). The hazard ratio associated with patient typing rate was 0.96 (95% CI = [0.93 ; 0.99]), $p$ = 0.019), suggesting that for every unit increase in typing rate (measured as words per minute) the likelihood of a patient dropping out of treatment was 4% lower, on average. Given the range of typing rates in the data set, this could be quite a large effect. Patient certainty (LIWC) and patient negation use (LIWC) were respectively associated with a hazard ratio of 1.28 (95%CI = [ 0.99 ; 1.66 ], $p$ = 0.058) and 1.27 (95% CI = [ 1.07 ; 1.50 ] $p$ = 0.005), suggesting, respectively, a 28% and 27% increase, on average, in the likelihood of dropping out for every additional percent of certainty or negation words used in a therapy session. The hazard ratio associated with patient guilt (expanded PANAS-X category) was much higher, 5.8 (95 %CI = [ 1.86 ; 18.08 ], $p$ = 0.002). This may be due to the low levels of guilt language picked up within the language in the transcripts as a one percentage difference in patient guilt language would be a large change.

Within therapist language, both measures of affect (expanded PANAS-X based I2E queries) were suggested to be associated with drop-out. However, both associations were in the same direction, suggesting that higher levels of both positive and negative therapist language had a protective effect against drop-out in this dataset. This may suggest that a therapist accessing and expressing affect of any kind was associated with lower likelihood of drop-out. However, the p-value attached to the association between therapist negative language and time to drop-out suggests only modest evidence for this effect, there is stronger evidence supporting the association between therapist positive language and time to drop out.

### 10.4.2 Validation set

The same process was followed to look at factors to explain drop-out in the validation data set. Data from 348 individuals were included in this analysis, with 146 failures (drop-out events) within this dataset. The time at risk was 11,797 days.

#### 10.4.2.1 Baseline model

As was the case in the previous dataset, the analysis was initially carried out with non-linguistic features.

**Table 10-8 Results of Cox Proportional Hazards model for drop-out from treatment from baseline measures**

| ***Covariates*** | ***Hazard Ratio*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***PHQ-9 score*** | 1.06 | [ 1.02 ; 1.09 ] | <0.001 |
| ***Number of sessions attended*** | 0.51 | [ 0.40 ; 0.65 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.11 | [ 0.67 ; 1.82 ] | 0.119 |
| ***Step group 3+*** | 0.43 | [ 0.16 ; 1.20 ] | |

The same set of predictors was associated with outcome in this model as was the case in the development dataset with similar hazard ratios associated with each variable included. A PHQ-9 score that was one point higher was associated with a 6% increased likelihood of dropping out of treatment, on average. Attendance to therapy sessions had an effect in the opposite direction with the results suggesting that for every session attended, the likelihood of dropping out of treatment was lowered by 49%, on average. The global test of proportional-hazards assumption, testing Schoenfeld's residuals, was not significant.

#### 10.4.2.2 Combined linguistic features model

This analysis was carried out on the same set of data with the inclusion of linguistic features as candidate predictors. These were originally tested in sets and the combined results will be presented here.

**Table 10-9 Results of Cox Proportional Hazards model for drop-out from treatment from baseline measures and linguistic features**

| *Covariates* | *Hazard Ratio* | *95% CI* | *P* |
|---|---|---|---|
| *PHQ-9 score* | 1.04 | [ 1.01 ; 1.07 ] | 0.009 |
| *Number of sessions attended* | 0.53 | [ 0.41 ; 0.67 ] | <0.001 |
| *Step group 2* | | | |
| *Step group 3* | 1.09 | [ 0.65 ; 1.82 ] | 0.149 |
| *Step group 3+* | 0.44 | [ 0.16 ; 1.24 ] | |
| *Agenda setting* | 3.99 | [ 1.59 ; 9.99 ] | 0.003 |
| *Patient social language (LIWC)* | 1.23 | [ 1.00 ; 1.50 ] | 0.048 |
| *Patient first person singular pronouns (LIWC)* | 1.15 | [ 1.06 ; 1.25 ] | 0.001 |
| *Patient positive language (LIWC-based I2E category)* | 1.03 | [ 0.99 ; 1.08 ] | 0.123 |
| *Therapist Positive language (Expanded PANAS-X-based I2E query)* | 0.61 | [ 0.38 ; 0.98 ] | 0.039 |

The results of the analysis suggest that the non-linguistic features were still significantly associated with outcome after the inclusion of a number of linguistic features. Step group was just statistically significant at the 15% level when the dummy variables were combined in a test of the overall categorical variable. PHQ-9 score and the number of sessions attended were, however, strong predictors with very similar hazard ratios associated as those presented in all three previous models of drop-out. In terms of linguistic features, five features were retained in this model. Three were patient features and two were features of therapist language. Within the patient language features were first person pronoun use (LIWC), social language (LIWC) and patient positive language (LIWC-based I2E query). All three hazard ratios associated with these features were above one, suggesting that higher levels of each of these linguistic features was associated with an increased likelihood of dropping out of therapy in this dataset. Taking the example of patient social language (LIWC), the associated hazard ratio was 1.23 (95% CI = [ 1.00 ; 1.50 ], $p$ =0.048). This suggests that for every percent of patient language in a therapy session that was qualified as social language in the LIWC, the associated likelihood of dropping out of treatment after that session was 23% higher, on average. This result associated with patient positive language was surprising as, given results throughout the models present, a protective effect of patient positive language against drop-out might be expected. However, the statistical evidence supporting this effect is very weak so further evidence would be needed to confidently report this association.

In the case of therapist language, positive language as measured by the PANAS-X based I2E query, was significantly associated with outcome in this model. This linguistic feature was the only one that was significantly associated with drop-out in both the analyses of drop-out in the development and validation datasets, with a similar associated hazard ratio, 0.61,( 95% CI = [ 0.38 ; 0.98 ] $p$= 0.039) in this data set and 0.64 (95% CI = [ 0.42 ; 0.98 ], $p$ =0.041). This suggests that for every percent of therapist language in a session that qualifies as positive language (PANAS-X based I2E query), the likelihood of a patient dropping out of treatment after that session was 39% lower, on average, in this data set.

The linguistic feature Agenda setting (CTS-R based) was also statistically significantly associated with drop-out in this analysis with a hazard ratio of 3.99, suggesting that references to agenda setting increased the likelihood of dropping out of treatment. As was the case with Guilt language in the previous dataset, the large hazard ratio associated with agenda setting may be due to the low levels of references to agenda setting picked up. However, the analysis suggests that more references to agenda setting increase an individual's likelihood to drop-out of treatment. If this effect is present consistently in the service provided, this would be a concern as agenda setting is an integral part of cognitive behaviour therapy.

### 10.4.3 Overview of results

The models of recovery presented here did not suggest there was high value in including linguistic features as predictors of recovery. In models of both PHQ-9 and GAD-7 based recovery, the linguistic features that were suggested to be predictive of outcome in the development set were not found to be significantly associated with outcome when tested in the validation set and the model performed poorly.

Considering the two sets of results of the survival analysis it is clear that there is little agreement on the linguistic features that are likely to influence drop-out, whether negatively or positively. Agreement across the models came with the influence of the PHQ-9 score, number of sessions attended and, to some extent, the severity of the mental health condition an individual was presenting with. The only linguistic feature that was suggested to be associated with drop-out across the two data sets, and with an almost identical hazard ratio, was therapist positive language measured by the PANAS-X based I2E query. This suggests that therapist positivity, or expression of positive emotion may play a role in keeping an individual in treatment.

# Chapter 11. Discussion

The discussion of the work reported on is set out in a number of stages. Initially, I discuss how text mining methods can be applied to the type of data studied here. This is followed by the interpretation and explanation of the results within the context of the project and how these relate to other relevant research work. A reminder of the research aims and design will follow, prior to a critical evaluation of these in light of the results and knowledge acquired throughout the project. The implications within both a clinical context and in terms of future research will also be discussed.

## 11.1 The application of text mining in online CBT

The main goal of this research was to explore how text mining methods could best be used when working with transcripts from online cognitive behaviour therapy. The answers gained from the work put into this project can be broken down into three aspects: the nature of the query development process, the selection and definition of linguistic features to focus on, and the testing of these.

### 11.1.1 Query development process

Firstly, it is important to understand the nature of the query development process for linguistic features, in this case within the I2E framework. It relies on the manual building of a query from what is essentially a blank page. Dictionaries or ontologies can be imported, which allows groups of terms to be inserted into a query. Previously built queries can also be incorporated into a new query. This process means two further challenges, one in relation to the reliance on human knowledge and skill to adequately develop a query – an idea that will be returned to in the next paragraph and the second is that the broader the construct the query aims to identify, the more difficult its development is likely to be. Accurately capturing expressions of a broad construct in text requires the researcher or query developer to identify all possible terms and phrases that refer to that construct. Furthermore, the broader a query and higher the number of terms referring to it, the more likely it is that ambiguous terms are included that then need to be qualified. Qualifying selected terms or phrases within a query in order to differentiate between their multiple uses or meanings comes down to extracting them from the broader list and defining them separately. For example, in the LIWC-based I2E query measuring positive language, the words 'like' and 'well' were leading to the inclusion of irrelevant results in cases where they were used as filler words. In order to remedy this, these two terms were not counted within the broad 'positive language' word class but counted separately with conditions attached to them, as a word within the utterance as opposed to at the beginning of the utterance. For example, in the case of 'well', uses of the word at the beginning of a phrase such as 'well, I think that…' were not considered to be positive phrases. Therefore, the word 'well' was removed from the dictionary so that is was not counted at every instance of the word and a new condition created so that it was counted only when it was not the first word in a phrase, such as 'it went well'. It therefore seems a more complex task to develop queries for broader constructs and focusing on smaller elements may be a good approach for future work. In addition, queries developed to focus on smaller, more specific constructs could potentially be combined into broader features when this is appropriate. Focusing on narrower features is likely to lead to more accurate queries.

### 11.1.2 Feature selection

The second aspect of text mining that can lead to an answer to this question is around the selection and definition of the features to extract from text. Query development itself requires the technical knowledge to work with the software and the linguistic knowledge to develop the parameters of the query skillfully. Prior to embarking on the task of query development, the features to be worked on need to be determined, selected and defined. This will ordinarily rely on both previous work carried out in the field or the involvement of experts and ideally, both. In this project, previous work pointed toward the LIWC dictionary as the primary approach to measuring particular features in language that were relevant to mental health. This was therefore a logical direction for the project to follow. However, from this point onwards, the path was not clear and a number of choices and, sometimes difficult, decisions had to be made to select a route for the project to take. For example, following the results from LIWC-based features, I had to decide whether affective and emotional language should be studied further from a different approach or whether the potential contribution of these features of language must be considered limited and not pursued further. It was decided that a new approach would be developed and tested. This aimed to remedy some of the concerns with the LIWC dictionary such as how broad it is and whether it is well adapted to the analysis of language used in a conversation. This then led to the use of the Positive and Negative Affect Scale (PANAS-X) and its expansion with the aim of being more applicable to the data set and providing a more narrow focus on affective language. Similarly, the selection and adaptation of items from the Revised Cognitive Therapy Scale (CTS-R) as linguistic features relied on the weighing up of information about the type of therapy provided by Ieso Digital Health and their use of the CTS-R in training and evaluation as well as background literature and an understanding of how well the features could be built into text mining queries in I2E. These are some of the challenges that were faced in this project and that have shaped my understanding of how best text mining can be applied with this context.

With this experience in mind and the knowledge that there is little work within mental health to support the development of text mining queries, I suggest an extra stage in this type of research in the future. Qualitative work would support the identification or adaptation of features of cognitive behaviour therapy and mental illness that could be measurable in therapy transcripts and useful to identify for both research and clinical purposes. In addition to selecting new elements to investigate, this kind of qualitative work would require clear definitions of features to identify. In the case of the CTS-R features for example, this process would be helpful in drilling down to the essence of the items and strictly defining what the query identifies. The aim would be to make these definitions as objective as possible as computerised analysis requires clear instruction. The involvement of mental health professionals and academic experts in query development alongside linguistic and text mining experts is also likely to create the ideal circumstances for the development of successful queries. As this research is so young and developing rapidly, it seems that collaboration between experts in different fields is the best path towards successful application of text mining to therapy transcripts and other textual data within mental health.

### 11.1.3 Feature validation

The final element to consider is appropriately testing the developed queries. This involves sensitivity analyses and relies on the previous point of clear definition of the features to be identified as well as the amount of context or length of the textual data with which a human rater is working. This decision about the document length could have a large impact on the result. Having access to the wider context of a phrase, over an entire therapy session, rather than just the words preceding and following it, can very much change its meaning. For example, if a patient is working on self-confidence and assertiveness during a therapy session and uses the phrase ‘I won’t do it!’ in reference to a demand that has been placed upon them, this could be interpreted in a number of ways. On the one hand, with knowledge of the rest of the session, it could be seen as an example of resolve, confidence and motivation in an individual. On the other hand, if this phrase is rated in isolation, it could be interpreted as negative or as evidence of conflict. This means that even sensitivity analysis results and the agreement between human raters will most likely be bound within the criteria set out for their undertaking. Such criteria might be whether utterances are judged individually or sequentially throughout a session transcript. Additionally, individuals who were not involved in the development of the queries or definition of the features to be identified would ideally complete sensitivity analyses to avoid bias. Overall, it seems that a lesson to be learned from this project is the necessity to employ as strict and clear a method as possible, something that can be quite a challenge when there is an element of subjectivity involved and features within language are context dependent.

## 11.2 Language features

### 11.2.1 Affect

The first set of linguistic features that will be discussed can be broadly referred to as affect. More specifically, these are the various measures of negative and positive language considered within this project. Three, arguably four, different approaches to measuring negative and positive language were applied. These were the LIWC categories of negative and positive language, the LIWC-based I2E queries of negative and positive language, the PANAS-X based queries of negative and positive language, and the subcategories (guilt, hostility, sadness, fear, joviality, self-assurance and attentiveness) within the expanded PANAS-X affective categories. However, at this stage in the discussion, these will be considered together as measures of affect and their association with outcome scores discussed generally.

Affective language and sentiment analysis (analysis of whether an individual is expressing a positive or negative attitude) is a much-researched topic within computational linguistics, with a vast range of approaches to sentiment or opinion mining, many of which rely on machine learning methods that go beyond the scope of this project. One major conclusion that has been drawn from reviewing sentiment analysis in different fields was the need for sentiment analysis tools to be adapted and customized to the field in which they are applied (Pang & Lee, 2008). Their value is very much context dependent and there are few tools that have been adapted for application within mental health and psychological therapy (Shickel et al., 2016). The dictionary measure of the LIWC has however been repeatedly applied within mental health research and negative emotional or sentimental language has often been associated with poorer mental health outcomes. A number of studies in different settings have put forward an association between levels of negative language and measures of depression or the presence of a depression diagnosis in an individual (Arntz et al., 2012; Molendijk et al., 2010; Rude et al., 2004; Tausczik & Pennebaker, 2010; Van der Zanden et al., 2014).

In every model predicting outcome developed there was at least one, if not more, measure of affective language included, whether this was within therapist or patient language. Measures of both positive and negative language were significantly associated with outcome during treatment in the majority of the models developed. In the models predicting outcome at the end of treatment, it was positive language (therapist or patient) that was significantly associated with outcome, whereas negative language was not. This suggests that though both negative and positive language may be associated with outcome scores throughout treatment, it seems that when it comes to predicting outcome at the end of treatment from language use early in therapy, it is positive language use that is important. Howes et al., (2014) found evidence of correlations between negative and positive language in session transcripts (patient and therapist language combined) and PHQ-9 scores associated with that session. Their results suggested that higher levels of negative language were associated with higher (worse) outcome scores and the opposite association was true of positive language. The results found in this project broadly support these conclusions as well as much of the other research work carried out in the field (Arntz et al., 2012; Howes et al., 2014; Park et al., 2013; Rude et al., 2004).

The same direction of association between affect and outcome score was found across all models, with positive language consistently associated with better outcome scores and negative language with worse outcomes. In previous work, the majority of the focus has been on the association between negative language and worse mental health outcomes. An interesting result was the association between positive language early in treatment and end of treatment outcome. There are a number of possible mechanisms behind this association. One possibility is that positivity in a therapist from the beginning of treatment encourages the patient to engage and boosts their confidence both in the therapy and themselves, thus making them more likely to succeed. Conversely, positive language from the therapist could arise as a response to the patient’s engagement and positivity, creating a circular effect. Therapist positivity may also be a marker of therapist confidence in their ability to work with a particular patient. These are also all factors that will most likely improve outcomes in treatment. Positive language early in treatment could therefore be considered either a marker of therapy potential or a feature that could be targeted in order to affect change, especially where therapist language is concerned.

Overall, the affect results suggest that there is clear evidence of a relationship between affect and outcome in the data considered in this project. Despite the small additional variation in the outcomes explained by these language features, their significance and the nature of the association is consistent across models and measurement methods. A primary difficulty in the interpretation and application of these results is in understanding the nature of the relationship between the expression of affect in both patient and therapist language and the outcome scores measured. Is this simply a measure of the affect expressed by the patient and the therapist in a therapy session? And if this is the case, will automatic measurement of affect in this way provide any information beyond what a therapist is aware of during the course of a treatment session?

Given the observational nature of the research, it is difficult to establish in the modelling undertaken whether there is a causal relation between linguistic features and outcome. The time between the measurement of the two variables in each association is the primary evidence for cause but this isn’t enough to rely on. Taking account of the similarities and differences in results across the different models (outcome before session, outcome before next session and outcome at the end of treatment) may also help in the interpretation of this relationship.

Overall, it seems that the affect measured in language may be a reflection of the patient’s mental health. This may be the case for patient negative language in particular, for which there were recurring significant associations with outcome scores recorded before a session. However, it is also possible that higher negative language is a product of the treatment session focusing on worse mental health outcomes.

Another recurring feature that was significantly associated with outcome was patient joviality. This was a measure of happiness and enthusiasm language. Unlike patient negative language, this feature was statistically significant in both models of outcome, namely before a session and before the next session. Patient joviality could therefore both be a marker of less depression or anxiety but may also reflect a patient's attitude towards and during treatment. A more positive attitude towards treatment could in turn lead to improved outcomes.

Alternately, if the level of affect expressed in a therapy session can improve mood and alter the mental health outcomes recorded this could be a potential mechanism for change, though this would primarily be possible through the modification of therapist language. Therapist positive language was associated with both outcomes before a session but only with PHQ-9 score before the next session. It may be that therapist positive language in the first two models is reflective of patient levels of depression and anxiety either by reflecting patient mood or due to the scores directly guiding either the content of a session or the focus of the therapist. In terms of the association between therapist positivity and PHQ-9 score at the next session, however, it may be that positivity in therapist language is encouraging to the patient or is evidence of a productive treatment session, which is turn may lead to improved outcomes. Depending on the nature of this relationship, there may be potential to adapt therapist language to give a patient the greatest chance of improvement. This is hypothetical, however, and needs further evidence. Further discussion of the differences between the models of outcome before session and outcome before the next session can be found in 11.3.2.

### 11.2.2 Non-affective LIWC features

In addition to the categories of negative and positive language, six further categories from the LIWC dictionary were tested across the models developed in this project. These were selected based on previous research on categories that were found to be associated with mental ill health and as features that may provide evidence of anxiety or depression in an individual's language. The six non-affective categories selected were negations, social language, first person pronouns singular and plural, insight language and certainty language. These were measured and tested in both therapist and patient language. When considering the LIWC linguistic features alone, the majority of these features were retained in at least one model of the six mental health outcomes considered. A smaller subset appears repeatedly in the models developed. These were levels of negations, either in patient or therapist language, and patient use of social language. The association between these features and outcome score (all versions) was positive, suggesting that higher levels of these language features used during a therapy session were associated with higher PHQ-9 and GAD-7 scores both before a session and before the following session, and therefore worse outcomes. In the case of negation use, this concurs with previous work by Arntz et al., (2012) who found that levels of negation use in individuals with a personality disorder following a course of psychotherapy for depression were higher at the beginning of treatment and lowered as patients improved and that levels of negation use became more similar with a non-clinical group over the course of treatment. Levels of negations have been seen to be higher in more emotional text (Pennebaker et al., 2001) and negation use has been regarded as evidence of a focus on what is lacking in an individual's life or what they are unable to do, which suggests a lack of need fulfilment (Arntz et al., 2012). It is possible that this is the case in the data studied here but it is also possible that negation use is evidence of a more closed or defensive position in the patient and possibly also the therapist, whether this is a consequence of the patient's style of expression or not.

The results associated with social language (including words such as ‘family’, ‘friends’, ‘talk’, ‘mate’) were less in line with previous work. Across the models developed, higher levels of patient social language were generally associated with worse outcome scores. This result differs from what would be expected based on other research work conducted using the same LIWC category. Previous work has found that more social language was associated with better adherence and attendance to treatment (Van der Zanden et al., 2014) and better scores on mental health measures (Cohn et al., 2004; Van der Zanden et al., 2014). This can been associated with theories of the protective effect of social contact and support against mental health problems, and in improving recovery rates (Silva, McKenzie, Harpham, & Huttly, 2005). However, in this project, the results appear to point to an opposite association with patients who use more social language recording worse outcome scores. One major difference between previous work and this project is the nature of the text being analysed. In previous work the textual data worked with has mainly consisted of personal narratives, whereas here it is conversational text within a course of psychotherapy in which a patient is likely to be expressing their difficulties and concerns and generally talking more about negative things. It is possible that social language used within this context is negative with social situations or circumstances even being put forward as a cause or trigger for problems. This may explain the opposite direction of association as if an individual is having difficulties with those who could provide them support and comfort, they are likely to record worse mental health outcomes.

Two related and surprising results were the lack of significance of measures of first person singular and plural pronoun use in the majority of the models developed. In previous work, first person plural pronoun use has been associated with improvement in mental health outcomes and better therapeutic processes (Haug et al., 2008). This feature was not statistically significant in a majority of the models developed here. It is however possible that the conversational nature of the textual data and the therapy format did not provide the same space for reflection on the sense of belonging to a group that personal narratives allow. Similarly, high levels of first person pronoun use in personal narratives have consistently been associated with low mood and mental health conditions such as depression and anxiety, reflecting the self-focus that often accompanies these mental health conditions (Mor & Winquist, 2002). However, the results in this project suggest that where first person singular pronoun use was significant, in models of PHQ-9 before a session and before the next session, as well as GAD-7 score before a session, it was negatively associated with outcome score suggesting that higher levels of first person singular pronouns were associated with better mental health outcomes. This therefore does not support previous work in this area. As has been discussed before, the conversational nature of the textual data may change how patients express themselves within this area and within psychological therapy sessions first person pronoun use may indicate greater engagement and that a patient is taking an active role in their treatment. Though this was not the aim of this research, there is potential for qualitative exploration of the relationships between language features and mental health outcomes. Here, the understanding of this is mostly speculative.

### 11.2.3 CTS-R features

The final set of linguistic features considered in this research project were measured through I2E queries based on the Revised Cognitive Therapy Scale. This is a scale used to rate therapist skill and adherence to the cognitive behaviour therapy structure and process. Four of the twelve items on the scale were selected for query development and testing within the model presented. Each of the language features developed was significant in one of the developed models, with agenda setting recurring as a significant predictor in models of PHQ-9 and GAD-7 score recorded before a session and of GAD-7 score before the following session. Agenda setting was also significant in these models when combined with significant features from other linguistic sets. In this last model (GAD-7 before the following session), homework setting was also significant. In all models predicting outcome measures the CTS-R features that were significant were negatively associated with outcome, suggesting, for example, that more references to agenda setting and homework by the therapist were associated with better mental health outcomes.

This supports the suggestion that adherence to the CBT structure leads to better therapy outcomes. However, previous work has had mixed results with some suggesting that greater therapist skill in keeping to the CBT structure is associated with better outcomes (Shaw et al., 1999) and others suggesting no difference in outcome whether or not therapists adhere to the CBT structure (Huppert, Barlow, Gorman, Shear, & Woods, 2006). In the analysis of drop out in the validation set, agenda setting was associated with a higher likelihood of drop out. This can be seen to go against the idea of references to agenda setting as a positive contribution to therapy but can also be seen as an indicator that some individuals dropping out of CBT may be doing so due to the structured nature of it. It is therefore not necessarily a discouraging factor for all patients but may help determine who will engage with and benefit most from CBT. It is also important to note that agenda setting was not significant in any of the regression models when they were re-calibrated with the validation data set. This may suggest that a different association between agenda setting and outcome is at play in the two data sets. The inconsistent results make it difficult to draw any firm or generalisable conclusions with regard to presence of agenda setting features in therapist language in online cognitive behaviour therapy.

Multiple studies have considered the impact of therapist factors on outcome in psychotherapy and have suggested effects of therapist empathy and experience (Luborsky et al., 1980) and interpersonal conflict resolution skills (T. Anderson, Ogles, Patterson, Lambert, & Vermeersch, 2009) as potential predictors of positive outcomes. These results are not undisputed, however, with some work arguing that college professors achieved similar results as highly trained psychotherapists (Strupp & Hadley, 1979) and much of the therapist ability is seen to reside in the ‘therapeutic alliance’ which has proven difficult to untangle (Baldwin, Wampold, & Imel, 2007).

In this research, interpersonal effectiveness, which is considered to be a key component in therapeutic alliance, did not appear as a strong predictor in any of the models. There could be a number of explanations for this, a key one being in the difficulty of accurately measuring such a difficult to grasp, and quantify, element of therapy. A more rigid definition and further development of the query in collaboration with therapists working online might provide more insight into how it can be measured in online therapy. The three other items considered were more structural and intended to be simpler to measure using a text-mining query. The rates associated with the measures suggested that the text mining queries measured low levels of each of these items in the transcripts. It is possible that this is a reflection of what is contained in the transcripts but it is also possible that references to agenda setting, homework and pacing in a session were missed by the queries developed if these were not broad enough. Indeed, the conclusions that can be drawn from these analyses are only as good as the queries on which they depend. In order to determine whether the queries are "good enough" a current gold standard is required. Currently, the CTS-R is a manual rating scale used by trained mental health professionals to evaluate the work of therapists in training and during supervision. Determining how well a computerised query performs would require manual examination of a number of transcripts to establish how many references to the feature studied were detected and how many were missed. This is a form of sensitivity analysis relevant to all the features studied that is further described in section 11.4.3.

To my knowledge, this is the first piece of work that has sought to use text-mining methods to identify evidence of adherence to these elements of the CTS-R, with perhaps the exception of Interpersonal effectiveness. The latter has been considered in part by work that has aimed to automatically measure empathy, such as that carried out by Xiao et al., (2015) who achieved an 85% accuracy rate in the classification of empathy in transcripts from motivational interviewing (Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015). Other work has looked into identifying specific elements of therapy in session transcripts such as identifying reflections (Atkins et al., 2014; Can et al., 2012) and identifying linguistic evidence of patients' motivation to change their behaviour (Tanana et al., 2015) in motivational interviewing. Success rates in performing these classification tasks were variable with work on reflections being much more successful than that predicting patient language expressing motivation to change. Though these results are not directly comparable to those presented here due to the method of analysis and context (Motivational Interviewing) within which they were applied, they may provide an avenue for further work. Machine learning techniques may be useful in identifying phrases that are evidence of the CTS-R items studied, especially where these were missed by human researchers, and further developing the text mining queries used to measure them.

## 11.3 Statistical modelling of mental health outcomes

### 11.3.1 Overview of results

The results presented in the previous eight chapters put forward multiple versions of models for the nine mental health outcomes. These included continuous and binary PHQ-9 and GAD-7-based outcomes associated with a given session, associated with the following session or reported at the end of treatment as well as a measure of survival (continuing treatment). With the exception of the model looking at time to drop-out, the developed models aimed to be predictive and the performance of the regression models with continuous outcomes was externally validated with a data set that was not used for linguistic feature development. Throughout all the developed models, it was clear that baseline PHQ-9 and GAD-7 scores were the strongest predictors of outcome scores both during and at the end of treatment. The majority of the overall variation in the outcomes explained across the models could be attributed to the relevant baseline score, a finding that is neither clinically nor statistically surprising. The focus in this project, however, was on the contribution of linguistic measures. Four sets of candidate linguistic measures were considered across the models presented: LIWC categories, LIWC-based I2E affect queries, expanded PANAS-X categories and PANAS-X based features, and CTS-R based features. With the exception of only two outcomes in the set of CTS-R based results; a group of one or more linguistic features in each set was retained in the models of every outcome considered. There was quite some variability in the statistical evidence supporting the reality of the associations measured, indicated by a number of higher p-values across the developed models, but a subset of predictors with seemingly strong statistical evidence behind them was nevertheless present across the models. Overall, this suggests that linguistic features were significantly associated with outcome and that there is some gain from including them in a predictive model. However, in most cases this effect appeared to be quite small and the clinical value of the additional variation in outcome explained is debatable. This contribution of the linguistic features tested to each of the models developed will be discussed in this section.

### 11.3.2 Mental health outcomes during treatment

In a first instance, I will consider the results of models looking to predict PHQ-9 and GAD-7 scores before a session and before the following session using linguistic data from each appointment attended. As mentioned above, the results presented suggested that in each set of linguistic features, a subset of the variables was significant. The CTR-S based features could be seen as the weaker set in terms of prediction, with no variables significantly associated with outcome in some of the presented models. Setting the CTS-R based features aside, it seems that the overall performance of the developed models was quite similar within the same outcome, whether these included the LIWC features, LIWC-based query features or PANAS-X based query features. The mean cross-validated R-squared and calibration slope were used as indicators of model performance. In the case of the mean R-squared, the additional variation in the outcome explained by linguistic features ranged from 2%, in the case of the model of GAD-7 score at the following session from LIWC-based I2E queries, to 4%, in the case of the model of PHQ-9 score before the session from PANAS-X based I2E queries. The contribution of the linguistic features appears to be very small. Additionally, in the case of the models developed with the CTS-R based linguistic features, the models developed appeared to be consistently 1-2% weaker, in terms of variance explained, than the equivalent models developed with the other sets of linguistic features. This does not seem like a large difference but in context of the additional variance explained, their contribution to the models appears to be almost half that of the other linguistic features tested.

Though overall the variation in outcome explained suggests useful predictive models, the focus in this project was on the predictive value of linguistic features that can be measured using text mining methods. The models developed with linguistic features tended to explain between 1 and 4% additional variance when compared to the same model developed using only baseline and demographic features. This supports the repeated and unsurprising finding that the baseline features are strong predictors of outcome but also suggests that, despite some significant associations with outcome, the linguistic features alone are not useful for inclusion in a prediction model for the PHQ-9 and GAD-7 outcome scores during the course of treatment. Put simply, the linguistic features were statistically but not clinically significant.

There are a number of ways to interpret these results that inevitably suggest further questions. The low additional variation in the data explained by the linguistic features in models of outcome score despite significant p-values suggest high variability in the data and therefore a large amount of error that the linguistic features included cannot explain. It may be that the individual words that a patient uses, or the expressed affect, measured within the categories defined in the Methods chapter, are not associated with the severity of a patient's mental health condition closely enough to provide strong predictive power of their recorded mental health outcome. This could be the case despite there being a general association between a number of individual linguistic features and outcome, or the presence of a mental health condition. This result does not stand against the research work that has shown differences in linguistic features between groups of psychiatric and non-psychiatric patients (Molendijk et al., 2010) or depressed and non-depressed individuals (Rude et al., 2004). In fact, the presence of statistically significant effects supports these to some extent, as there was evidence of associations between various linguistic features and outcome scores. The task at hand here was, however, a little more complex as the aim was not to discriminate between clinical and non-clinical groups but to predict continuous outcome scores within a clinical population.

Despite statistical significance, the results presented do suggest that the nature of these associations is not strong or consistent enough to provide a great deal of predictive power. This explanation would not discount all language features within a course of psychotherapy as predictors of outcome but suggests that the linguistic features investigated in this project had limited success in this task. It is still possible that different linguistic features, not investigated here or in other work looking at similar data, would provide stronger predictors of outcome. Certain features such as Guilt as measured by the PANAS-X based queries and homework setting (CTS-R based) did not have high prevalence throughout the transcripts, a factor that will weaken the measurement of an association between these and outcome measures. It is possible that future research will put forward more successful predictors of outcome. Future research directions will be discussed at the end of this chapter.

Additionally, the severity of a mental health condition may affect the type of association between language features and reported outcomes. We could speculate about mechanisms that could be at play here. The population included in this study was made up of individuals with mild to moderate depression and anxiety. On the one hand it is possible that any associations between linguistic features and depression and anxiety scores would be stronger and clearer in a more severely affected population meaning that, with a less severe population as was the case here, the association were less clear. It is also possible that the nature of the association differs between individuals. A person who has a severe affective disorder may find themselves unable to control their emotions and therefore express themselves in highly affective terms, for example, but another individual may find that they are unable express that emotion verbally or that they have or are disengaged from it. This would go against the idea of a more measurable association between language use and outcome being found in a more severely affected population. These are two of multiple possible circumstances that are likely to affect the relationship between an individual's language use in therapy and their mental health outcomes, and which may have had an impact on the results found in this project.

A second possible interpretation of the low variation in outcome explained by the linguistic features tested within these models is that the measures used or developed are not appropriately measuring what they intend to measure. Construct validity may be a concern here. If the measures developed are not accurately measuring a given linguistic feature, it may mean that, for example, one way of expressing an idea or feeling was being consistently missed. This would in turn mean that different expressions of the same feature would not have been measured consistently and therefore make statistical evidence of an association less likely. The low amount of variation in outcome explained by the linguistic features could therefore be down to the method of measurement and not a lack of relevance of a given construct in predicting outcome. A good number of research studies have assessed the validity of LIWC categories as measures of affective and emotional language (Kahn, Tobin, Massey, & Anderson, 2007; Tausczik & Pennebaker, 2010) and there is some evidence supporting the validity of the PANAS-X as a measure of emotion (Watson et al., 1988). However, there was no specific validity testing carried out for the queries developed throughout this project. Furthermore, LIWC analysis has primarily been carried out within the context of self-narratives, as opposed to conversational data, which may weaken how well the validity results can be applied to this data set. The queries used in this project were developed using an iterative process that relied on manually checking results and editing a query when errors were apparent, but this does not exclude the potential for errors or omissions in development. It would be difficult to identify a consistently missed phrasing of a feature without extensive qualitative analysis, for example.

A third and important idea within the discussion of these results is the influence of individual differences on both outcome score and the way an individual expresses him or herself verbally. It seems logical to consider that there are vast variations in the way individuals presenting for psychological therapy will express themselves. Educational background, cultural background and personality are all likely to affect both an individual's relationship with their mental health condition and the type of language they use to express this. Furthermore, the level of emotional disclosure and way individuals choose to speak, or type, about their mental health condition is likely to vary greatly. This variation within the population may make it difficult to ascertain the size and significance of associations between linguistic features and outcome scores or even the direction of these as it is feasible that opposing directions of association are present across the population. It may be that very large sample sizes are necessary if interactions between multiple personality factors and linguistic features are to be considered. In a case of depression for example, an individual's language may become more detached and less emotional or, on the contrary, much more emotional and display visible distress. Taking into account differences in personality, education and culture and the interactions of these with mental health measures may lead to improved predictive models.

###### *11.3.2.1.1.1 Relevance of time of outcome measure*

The mixed effect models developed throughout the previous chapters were developed to consider psychological status at two different time points as measured by the PHQ-9 and GAD-7. The first considered the score recorded prior to a therapy session and the other, the score recorded prior to the following session. This allowed two, potentially different, associations to be considered. The first more cross-sectional with a shorter distance in time between the language analyses and the mental health outcome score recorded, perhaps meaning that the outcome score reflects the tone of the session better and the second with a greater lag between language use and outcome measurement with greater potential for the language measured to predict change in mental health outcomes. Throughout the sets of linguistic features tested, the estimated additional variation in outcome explained by linguistic features was consistently slightly stronger for models predicting outcome score prior to the session than for models predicting outcome score prior to the following session.

Most research using the LIWC for linguistic analysis works on the assumption that language use represents individual's internal world and mental state (Pennebaker et al., 2003a; Tausczik & Pennebaker, 2010). If these results are interpreted within this context, the weaker models predicting outcome at the following session could be seen as a product of the time gap between the production of the language used for analysis and the recording of the outcome variables considered. In this case we might expect to see the same set of predictor variables in the two models as well as these being more focused on the language used by the patient, as it is their outcome score being considered. This doesn't seem to be the case in the results presented. Let us consider the models developed by combining the significant candidate predictors from the individual sets of linguistic features, presented in Chapter 9 as an example. It appears that though there was some overlap in the linguistic features that were significant in the models at the two different time points (outcomes 1 and 3, and outcomes 2 and 4), there are also a number of differences. In the results in Chapter 8 (Table 8-1), predicting PHQ-9 score recorded before the session in all cases in the dataset, six patient language features and three therapist language features were significant. In the equivalent model fitted to predict PHQ-9 score before the next session, only three of these were significant, of which only one was a feature of patient language. A similar pattern was present in the GAD-7 versions of these models with only one patient language variable being significantly associated with outcome in the model predicting outcome score at the following session. It may therefore be that with the models looking at a future outcome score, the results were illustrating associations between therapist language use and changes in outcome score before the next appointment, associations that could be indicative of therapeutic processes in action. The individual linguistic features were discussed in 11.2 but generally, the presence of this kind of association and the capacity to measure it could provide important insight into which elements of a treatment session impact on a patient or trigger change to the extent that this is reflected in their mental health outcomes measured before a future session.

### 11.3.3 Models predicting end of treatment outcome score

A different set of regression models was developed to consider the association between language use early in treatment and outcome scores reported at the end of treatment. The aim here was to determine if there are features of language that occur in the first two treatment sessions that may provide an indication of how successful treatment will be for a given individual. Where linguistic features were significant in these models the additional variation in outcome explained ranged between 6% and 13%. At the lower end, the additional variation in outcome explained was not large but nonetheless significant and at the upper end, 13% is a considerable improvement on a model that previously explained 20% of the variation in end of treatment GAD-7 scores when only baseline measures were included. In the models developed by combining the previously significant linguistic features in each set (Chapter 9 – Outcomes 5 and 6) the additional variance explained by these features was 9% in the PHQ-9 model and 13% in the GAD-7 model. These results suggest that there was a clear association between the use of a set of linguistic features (patient and therapist positive language, patient and therapist use of negations and patient social language) in the first two treatment sessions and the final outcome score recorded by patients. Though the same linguistic features appear in a number of the mixed effects regression models, the additional variance explained in end of treatment regression models is clearly greater than that in the mixed effects models. It appears that the language features at the beginning of treatment have greater predictive power for end of treatment outcomes than language use at a session for the outcomes measures recorded before that or the next session.

A possible explanation for the difference in explained variation in outcome is that the relationship being measured was qualitatively different in the two types of models. In the mixed effects model the association may be more about an individual's mental state expressed in their language use or the impact of one therapy session on this in the days following. The models predicting end of treatment outcome scores may however be looking at features of language early in treatment that suggest engagement by the patient and therapist or other elements that are setting the treatment up for success or failure.

It is also possible that the first two treatment sessions, not including the assessment session, provide a slightly different type of conversation than sessions later in therapy. In the early treatment sessions there may be more conversation about understanding both the problems a patient presents, as well as how cognitive behaviour therapy works and what can be expected in the sessions that follow. Following the formulation and understanding of a patient's condition and circumstances in early sessions, later sessions may involve more checking in with patients about progress with homework and goals and therefore less emotional or expressive language. Emotional language is what has primarily been focused on in previous research associating LIWC categories with mental health (Arntz et al., 2012; Cohn et al., 2004; D'Andrea et al., 2012; Molendijk et al., 2010) However, these were primarily in personal narratives as opposed to sessions of cognitive behaviour therapy in which the language may be more regulated due to the awareness of a reader, and even less emotional during the more goal-oriented sessions.

### 11.3.4 Performance of models on an independent data set

The predictive models discussed above were subsequently validated through the application of the prediction models to a new dataset and the assessment of how these models fared. There were a number of differences between the populations in the development and data set. There was a geographical difference in the NHS trusts under which the patients were being treated as well as approximately a year's difference in the time of treatment. Within that time a number of improvements had been made to service provision provided by Ieso Digital Health, including therapists being trained not to be overly familiar with patients or break into shorthand and common abbreviations when typing. Finally, in the development data set a large number of patients had been on a waiting list for face-to-face CBT for up to a year prior to being referred for treatment with Ieso, whereas this was not the case in the validation set. This information was supplied by Ieso but specific details of how long a patient had been waiting prior to referral to Ieso are not available within this data set. The mindset with which therapy was entered into is therefore likely to have been different between the two populations. Mindset here refers to the attitude the patient may have towards the offered treatment as well as their condition. This will include their belief in the value of the treatment and their trust in the service (and attached therapists), for example. If an individual has been on a waiting list for up to 12 months, it is possible that their response to their mental health difficulties has evolved (developing better coping strategies perhaps) as compared to an individual who is able to access treatment swiftly. Spontaneous recovery during this waiting time is also possible. Similarly, a long waiting list may foster frustration or even disillusionment in a patient population, thus potentially negatively affecting the attitude held towards the treatment when it does start.

The results of validation testing were presented in Chapter 9. The performance of the mixed effects and linear regression models was judged based on two statistical measures and graphical observation of mean predicted and observed outcome scores. The statistical measures applied here were R-squared, to estimate the variation in the outcome in the validation data explained by the developed model and the calibration slope, the slope of a regression model in which the predicted values are the sole predictor of the observed values. There was a pattern in these results across the mixed effects regression models. The calibration slopes estimated were good (0.90 and above with 1 indicating a model with good calibration providing accurate predictions on average) when estimated using a simple linear regression equation but slightly stronger when estimated using a random intercept model. Including patient identity as a random effect allowed the intercept to vary by individual and therefore account for clustering in the data (more similar scores within a patient than between patients). Furthermore, the same pattern observed in previous results was evident. Namely, that the model predicting PHQ-9 score before a session was the best calibrated, followed in turn by models predicting PHQ-9 score before the next session, GAD-7 before a session, and GAD-7 before the next session.

The results were not so promising when the variation in outcome explained by the baseline variables alone was taken into account. In the external validation, it seems that there was little to no gain in variation in the data explained from the inclusion of the linguistic variables, suggesting that these did not improve model fit. With the small gains provided by the inclusion of the linguistic features in the developed models it may not seem surprising that these were almost non-existent when the model was tested on the external data set. Additionally, there was clear evidence of differences between the populations in terms of the associations between language use and mental health outcomes. In each model validated, a number of the linguistic features that were significantly associated with outcome in the development set were not so when the model was recalibrated, or refitted, on the data from the validation set. The presence of irrelevant variables in the models is likely to have increased the error in the predicted values.

Together, these results suggest that though validation of the models was reasonable with models that held up and were predictive in a new data set, this was primarily due to the contribution of the baseline variables in the model with little to no apparent value for the inclusion of the linguistic features considered. Though a subset of the linguistic features were maintained as significant predictors in the models when re-calibrated using the external dataset, these did not appear to contribute much additional explained variation in outcome scores. These results suggest that in this format, the linguistic features studied do not add enough to the models of outcome to be useful in clinical practice. However, the service may continue the study of different or redefined linguistic features in order to improve model results with a large data set and/or the inclusion of other linguistic features.

The results for the linear regression models predicting end of treatment outcomes were promising for the model of PHQ-9 score with the model explaining 28% of the variation in the data in the validation set. The associated calibration slope was unfortunately weaker than those presented above and the scatter plot of observed and predicted values also showed quite a wide spread. The pattern was similar for the model predicting GAD-7 score at the end of treatment but with a weaker R-squared, with only 16% of the variation in the outcome in the validation dataset explained by the model fitted on the development data. The amount of additional variation explained in the validation data set by the inclusion of linguistic features was, as was the case in the mixed effects models, minimal to non-existent. The results of model re-calibration support and explain these results as only baseline scores and use of negations were maintained as significant predictors of outcome when the model was refitted on the validation data. These results suggest that the models of end of treatment outcome from language use early in treatment did not transfer well to a new dataset. This is not to say that the models should be discarded as there was still a substantial amount of variation in outcome explained but the value of including the linguistic features is limited. It may however be the case that this form of model should be adapted and re-estimated in different populations to provide more relevant outcome predictions.

### 11.3.5 Clinical outcomes

#### 11.3.5.1 End of treatment recovery

The results of analyses of two sets of clinically relevant outcomes were presented in Chapter 10 of this thesis. These were recovery at the end of a course of treatment and drop out. The analyses of recovery were based on a definition within the IAPT service of recovery as being achieved by an individual who reports a PHQ-9 score below 10 and a GAD-7 score below 8 as these are the thresholds above which it is suggested that an individual would benefit from psychological treatment. The models developed used measures of language features early in treatment as candidate predictors and were estimated to have a c-statistic of 0.82 (95% CI = [0.76 ; 0.88]) for the model predicting PHQ-9 based recovery and 0.81 (95% CI = [0.75 ; 0.87]) for the model predicting GAD-7 based recovery. This was a gain of 0.04 and 0.07, respectively, as compared to the equivalent models including only baseline values of PHQ-9or GAD-7. These are reasonably strong predictors of recovery and the inclusion of linguistic features improved the model in the development data. However, as was the case in the other models presented, when these models were tested on the validation data set, the gains of the model including linguistic features over the baseline model disappeared. These results are, understandably, in line with those suggested by the linear regression prediction of end of treatment outcome scores but provide a more practical measure for the service provider.

The binary definition of recovery used here is one used throughout IAPT practice and one that is often used in the evaluation of services and consequently in determining future resource allocation. Two recent pieces of work developed logistic regression models of recovery with varying success. The first, included in a report on the results of the first year of the IAPT initiative suggested a model of recovery based on a range of demographic, site and baseline outcome scores that was able to determine recovery accurately in 67% of cases (Gyani, Shafran, Layard, & Clark, 2013). The second piece of work was more successful in predicting a positive or negative clinical outcome (on the same criteria) from predictors such as gender, ethnicity, self or GP referral, baseline score, a measure of deprivation and English language proficiency. The model was able to correctly predict a positive outcome with 69% accuracy (overall percentage of correct predictions) and a negative outcome with 79% accuracy (Green et al., 2015). The results presented here sit broadly in line with those presented by Green et al. (2015) and given the differences in the set of predictors put forward it is possible that combining these would further improve the accuracy of recovery prediction.

The most closely comparable piece of research to the result in this project is the work involving classification experiments carried out by Howes & Purver (2014) on a subset of 882 online therapy sessions from the development dataset studied here. They looked to classify outcome scores by whether they were above or below the recovery threshold (PHQ-9<10) but this was done for each session rather than for each patient. They used various combinations of affect, high-level or baseline features (e.g. number of words used, patient gender, patient age) and extracted topics (clusters of co-occurring terms) within a session transcript to predict the binary outcome associated with that session. The results reported in chapter 11 appear to perform better than those reported by Howes & Purver (2014) but the two results are not directly comparable for a number of reasons. In terms of method the results in this project considered language use early in treatment as a predictor for end of treatment outcome score as opposed to the language within a session for its associated outcome score as was the case in Howes & Purver (2014). The F-score was their chosen reporting statistic, whereas in this project the c-statistic was used and these scores are not directly comparable though they are different ways of expressing the success of a given model. The F-score is calculated from measures of precision (true positives within all identified) and recall (number of positives retrieved over all present in the data), whereas the c-statistic is based on the probability that a random individual who experienced an outcome will have a higher predicted probability of experiencing the outcome than a random individual who did not experience it (Austin & Steyerberg, 2012). Though it is difficult to compare the models developed directly, their success does seem within the same range when evaluated within the data they were developed on. Within one data set these results appeared to support the idea that language used in a therapy session may provide additional information over and above baseline variables as to whether a patient is likely to recover or is in recovery. However, the difficulty is in defining language features that will generalize across populations as the value of including language features in the predictive models developed appears to be lost when tested on a new data set.

#### 11.3.5.2 Drop-out

The second clinically relevant outcome considered in the analyses in chapter 11 was drop out. Analysis of drop out was conducted to explore associations between levels of linguistic features in session transcripts and likelihood of dropping out from treatment. Cox models were used separately in both data sets to consider any linguistic features that were potentially associated with drop out. In each data set, the results suggested that a number of the linguistic features considered were significantly associated with drop-out, though only one of these features was present in both models. There has been some previous work looking to understand and predict adherence to treatment within psychological therapy but only limited work has looked at specific word features. Howes et al., (2012) developed a unigram-based model in which model parameters were devised by machine learning methods based on the association between patterns of individual words and an outcome, in this case high or low adherence. The results were promising; with the model achieving over 90% accuracy (overall correct classification) in predicting adherence as rated by the clinician but the machine learning nature of the model makes the factors difficult to interpret (Howes, Purver, McCabe, Healey, & Lavelle, 2012b). In a secondary analysis of collected data, another piece of research carried out within an IAPT service applied logistic regression to study the association between session attendance and a number of demographic variables, illness length and baseline scores. Their results suggested that a higher frequency of thoughts such as ‘I would be better off dead’ and of self-harm were associated with higher rates of non-attendance (Di Bona, Saxon, Barkham, Dent-Brown, & Parry, 2014). Though there was very little linguistic analysis in the aforementioned study, it does support the idea that higher severity of mental distress can increase likelihood of drop out.

In the development set, four patient language features were retained in the model of time to drop out. These were patient typing rate, patient certainty, patient negation use and patient guilt. Patient typing rate was the only protective factor and can be interpreted to suggest that patients who typed more during their therapy sessions may have been more engaged in treatment and therefore more likely to continue treatment, though factors such as education may also play a role here. This result supports that found by Van der Zanden et al. (2014) in a study of patients completing an online course of psychological treatment, in which patients who wrote more on the initial application form for treatment were found to better adhere to treatment (Van der Zanden et al., 2014). The three other patient linguistic features mentioned above were suggested to increase likelihood of drop out. Patient use of negations was previously associated with worse outcome scores during treatment. This may suggest either that patients with worse outcomes were dropping out of treatment, a result supported by the significance of the PHQ-9 score variable in this model and/or that these language features were an indication of non-engagement in this population that then lead to worse outcome scores. In both cases, higher levels of negations, with words such as ‘can’t’ or ‘don’t’, may suggest a more negative or non-engaging mindset in the patient at the time. Higher levels of patient certainty and patient guilt were also associated with higher drop out. As mentioned in the Results section, the low rates of guilt language are likely to be responsible for the high coefficient associated with this factor. As rates of guilt language are very low with a narrow range of values, a unit change of 1 (1%) represents a greater and rarer difference than, for example, a one-unit change in negative language use. The associated odds coefficient was therefore higher to account for the narrower range of values. In the cases of both guilt and certainty language it is difficult to speculate on the mechanism behind the association without some level of qualitative analysis of the phrases and patients in question. Certainty may be associated with drop out when patients are convinced that they should not or cannot continue treatment and guilt with a feeling that they are not worthy to receive treatment. Depression is often associated with feelings of worthlessness (McKenzie, Clarke, Forbes, & Sim, 2010) but the extension made here to being worthy of treatment is only a suggestion about the nature of the relationship between the language features and drop out rates.

In this same model, therapist positive and negative language as measured by the PANAS-X based queries were both associated with outcome in the same direction but the statistical evidence supporting the association with negative language was much weaker than for therapist positive language. Nonetheless, this was a surprising result as we may have expected negative language from the therapist to be associated with a higher drop out rate rather than a lower one. However, it may be the case that the higher levels of emotion and affective in therapist language, regardless of specific valence, suggest more engagement on the part of the therapist and potentially an improved relationship between the therapist and patient. Looking further into the affect expressed in therapist language and perhaps determining whether this is associated with a therapist reflecting back or clarifying patient language would allow further conclusions to be drawn about the context of therapist affect and the mechanisms that might be at play in its association with outcome.

In the validation set, five linguistic features were suggested to be associated with drop out. Only one of these was also significantly associated with outcome in the development set: therapist positive language as measured by the PANAS-X based query. The associated hazard ratio was also almost identical in the two models, with a protective effect against drop out of higher levels of therapist positive language. Of all the linguistic features considered this may be one to be aware of in the future as its association with drop out is not necessarily population specific. As mentioned previously, higher levels of affect may indicate a better therapeutic relationship and engagement of the therapist with the patient. It is also possible that higher levels of therapist positive language are encouraging to a patient and make treatment more pleasant, making them more likely to return. It may also be an indication that the therapist feels the treatment is going well and of their confidence and competence with a particular patient, which in turn would be associated with an individual’s likelihood of adhering to treatment. Any of these mechanisms is also likely to impact the therapeutic alliance between a patient and therapist, a factor that has repeatedly been put forward as a predictor of therapy outcome (Ardito & Rabellino, 2011; Baldwin, Wampold, & Imel, 2007). Therapist positive language was also significantly associated with outcome in logistic and linear regression models predicting outcome at the end of treatment from language used early in treatment. Though the measurement method (LIWC, LIWC-based or PANAS-X based) was not necessarily consistent, the underlying feature being measured was therapist positive language. The causal direction of the effect is not clear but the significance of therapist positive language across outcome and drop out models suggests that this linguistic feature may play a very important role in both patient outcomes and adherence to treatment.

Beyond therapist positive language, the only other therapist language feature significantly associated with outcome in this model was agenda setting. As with patient guilt in the previous model, the high coefficient was likely to be associated with the low rates of agenda setting language across the data set. However, more references to agenda setting were suggested to be associated with a higher drop out rate in this data set. As mentioned in the previous chapter, this may be a cause for concern for the service as agenda setting is an important part of CBT. As with previous linguistic features, it is not clear in what context these references to agenda setting were being made. On the one hand, it is possible that high rates were found in sessions in which a patient may have needed reminding about the agenda or bringing back to the agenda, which may in itself be an indication of low engagement from the patient. On the other hand, it is also possible that repeated references to agenda setting and a more rigid approach to the treatment session were a cause for irritation or disconnect in patients.

Some recent qualitatively focused work provides some context and contrast to these results. Ekberg et al. (2015) carried out a combined qualitative and quantitative analysis of the assessment session and length of treatment in transcript data from the IPCRESS trial (D. Kessler et al., 2009), the effectiveness trial for the online CBT used in this project. Their results suggested that when therapists provided more information about what would happen during a session and the rest of treatment (called ‘expectation management’ in the report), patients remained in treatment an average of 1.4 sessions longer (Ekberg et al., 2015). Agenda setting was a part of ‘expectation management’ in this work, suggesting that these results are not quite in line with those presented here. Further work in this area is both needed and would be a good candidate for a collaboration of qualitative analysis and text mining research. Forms of ‘expectation management’ could be operationalised into text mining queries and thus allow analyses of a larger population sample.

Two patient features were significant in the analysis of drop out in the validation set. These were social language and patient use of first person singular pronouns. Patient social language and patient first person singular pronoun use were both previously associated with worse mental health outcomes in this project and first person pronoun use has also been associated with mental ill health in a number of previous studies (Arntz et al., 2012; Consedine, Krivoshekova, & Magai, 2012; Haug, Strauss, Gallas, & Kordy, 2008). Higher levels of social language at application to an online psychological therapy course were associated with higher levels of adherence in previous work (Van der Zanden et al., 2014), an opposite effect to that suggested by these results. However, throughout the other models, social language was associated with a worse outcome, which appeared to be a risk factor for drop out.

Finally, patient positive language was also associated with higher likelihood of drop out but with only weak evidence supporting the effect and the associated hazard ratio was low, suggesting a small increase in likelihood of drop out. This result also goes against the idea that features generally associated with worse outcome increased the likelihood of drop out due to less engagement and progress in treatment. With positive language, it is possible that the opposite mechanism is at work in that patients who felt they did not need CBT or were coping quite well without it were using higher levels of positive language when expressing themselves. Drop-out from treatment has been documented in both individuals who feel they have improved and those who see little improvement (Bados, Balaguer, & Saldaña, 2007). For the features described in this section, qualitative analysis of the therapy sessions of individuals who dropped out is likely to help tease apart and understand these associations.

## 11.4 Were research aims met?

### 11.4.1 Research aims

The overall aim was to explore the potential of text mining in the analysis of online cognitive behaviour therapy for both research work and service provision. The first objective was to understand which linguistic features might be most useful. This involved selecting three sources of linguistic features that were developed and applied using text mining software in order to understand how these methods might be best applied to the data at hand. A series of statistical models was then developed in order to understand the impact of patient and therapist language features on outcome measures in a predictive model, as well as any association between language use and likelihood of drop-out.

### 11.4.2 Reminder of methodology

This project was an exploration into the method of text mining and how it could be applied to transcripts from online text-based cognitive behaviour therapy. Linguistic analysis is a growing field across a number of disciplines including healthcare. Vast amounts of personal and healthcare data are now being recorded, both digitally and in various textual formats, and it is not yet clear how best these data can be used. This project emerged as a partnership between UCL and two commercial enterprises, one of which had a specific therapy data set it was looking to investigate – Ieso Digital Health Ltd. - and the other, a method with which to do so; text mining – Linguamatics Ltd. The broad shape of the project was therefore very much affected by the partner companies involved. They also influenced and assisted the project throughout. For example, as text mining specialists and providers of the software used, Linguamatics played a crucial role and informed the query development process and the various stages this followed. Additionally, during discussions with Ieso about the development of CTS-R based queries, it became apparent that automatically extracting these features could be a step towards assisting therapist supervision and therefore help the company to manage their growing demand. Thus Ieso were supportive of the idea put forward. The involvement of both companies was therefore primarily tangible in the broad approach followed but their input was valuable and affected decisions throughout the research process.

Text mining had never previously been applied to this therapeutic data (indeed very few linguistic analysis methods had) and it is a method that revolves around a researcher building queries with which to interrogate the text. This meant that there was a need to know what was being searched for prior to analysis. This is what led to the application and then further development of the LIWC as this was the primary form of linguistic analysis that had previously been applied in mental health research. After achieving significant but limited results with LIWC and LIWC-based affective measures, two different sets of features were considered, both based on manual assessment scales. The first was a different approach to measuring affect, with a more restricted dictionary to contrast the large LIWC categories, and the second a set of items selected from a scale developed to rate therapist skill and adherence to the CBT structure. Each feature was selected on the basis that the construct they aimed to measure was seen to be associated with mental health outcomes and that if these could be measured and associated with outcome scores in online therapy, there would be scope for monitoring therapy progress and improvement of service provision through more individualized care.

The statistical analysis carried out to determine how features measured using text mining methods were associated with mental outcomes mainly involved regression models. The idea was to develop models using linguistic measures to estimate an individual's current psychological status as well as his or her future outcomes in this treatment format. These could then potentially act as a form of second opinion to the therapist in future. More accurate outcome prediction based on language use would allow for a more personalized approach in that action could be taken if, for example, it became clear that prognosis in therapy was not positive.

Using feedback from questionnaire-based outcome measures is an approach that has been the subject of previous research, Lambert and colleagues have tested this approach in a university counselling centre. In an experimental group they provided feedback to therapists about patient prognosis using scores from outcome questionnaires completed by patients before a therapy session as guide to patient progress. They found that feeding back to therapists led to improved recovery rates in patients who had a poor prognosis at a midway point as compared to the patients whose therapists did not receive feedback. There was no difference when the prognosis at this point was good. Though the improvement was primarily associated with patients then having a higher number of therapy sessions, it was nonetheless an improvement in outcome associated with a more personalized approach (Lambert et al., 2001) The same research group found the same effect across a variety of mental health conditions and treatment formats (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005; Probst et al., 2013) .

Recovery rates associated with cognitive behaviour therapy within IAPT are variable across services but the mean recovery rate across England is approximately 45% (Community and Mental Health team, Health and Social Care Information Centre, 2015). This means that if it is possible to provide extra support for, or refer to a different treatment approach, individuals who are unlikely to be successful with CBT therapy, this would be a worthwhile approach to take. Predictive models based on language use were a potential method of pre-empting bad therapy outcomes. Similarly, drop out rates from treatment seem to be quite high in IAPT services. They have been difficult to estimate as definitions vary across services. A report of IAPT provision in the first year it was rolled out suggested 38% of patients completed their course of treatment with 22% dropping out and a further 20% not completing for unclear reasons. The remaining patients had either declined treatment or been deemed unsuitable for CBT (Glover, Webb, & Evison, 2010). An exploration of drop out was therefore also included in this project so as to explore any factors that might explain drop out rates and therefore provide an indication to the service of how action might be taken to reduce these.

### 11.4.3 To what extent have these aims been reached?

Addressing these aims directly is best achieved by summarising the Discussion in sections 11.2 and 11.3. The pattern of results for mixed effects models predicting outcome throughout the course of therapy, was broadly similar. Though a number of the linguistic features developed and measured were significant in these models, their contribution in terms of additional variation in outcome explained was generally very small, making their importance in the model questionable. There was a clear statistical association between measures of affective language and negations in language use in therapy (details in section 11.2.2) and outcome scores, which is maintained during external validation of the models developed. The role for other linguistic features tested is less clear such as those based on the CTS-R scale and some categories of the LIWC that have been less consistently associated with mental health outcomes in this and previous work. However, as mentioned previously, the contribution of the linguistic features in predicting outcome was low and even negligible as observed in external validation, making their clinical value as predictors debatable. Their inclusion in routine practice is likely to depend on further development of the measures in the hope of more powerful and reliable results as well as the cost of routinely measuring these features. This is an idea that will be returned to in considering the implications for service provision of this research (in section 11.5).

The most promising model in development was that predicting final outcome from language use early in treatment as it suggested a greater amount of additional variation in outcome explained by a number of language features, in particular negation use and positive language use. However, with the exception of negation use, none of the linguistic features that were significantly associated with outcome in the development set model remained so in the validation set. This suggests that despite this model being promising in the development stage it lacks validation and does not seem generalisable across other populations. Broadly speaking, the linguistic features measured have provided some interesting information but with limited application at present. It is likely that the selection of features and measurement of those selected require much more work before routine application to research or practice could be envisaged.

The work carried out in this project has led to a number of conclusions regarding the potential of text mining as an analysis method and how best it can be applied within this therapy format (details in 11.1). Most importantly, the process of query development is a manual process. It seems that smaller, more specific pieces of information might be more suitable for identification and extraction using text mining methods than broader word categories. Identifying which specific elements and how best to define and build queries to extract these is likely to require extensive reflection for each feature considered. This should involve qualitative work looking at transcripts from therapy as well as the involvement of a multi-disciplinary team including mental health professionals, researchers in linguistics, dialogue and interaction as well as text mining experts (who also often have a background in linguistics) in order to identify which elements are both objectively measurable and of interest in research and practice, how these might be expressed in treatment, and how best to measure these.

Finally, sensitivity analyses of the features should ideally be carried out to test the queries with the assistance of a team of independent raters not involved in the development process. This would take the form of a sensitivity analysis to assess the performance of computerised queries in comparison with manual coding, aiming to assess how good the developed queries are at extracting what they were designed to extract. In order to complete this task, clear definitions of the individual features to be coded would need to have been devised, ideally by a team of both mental health professionals and experts in language and interaction (as above) and prior to query development so that all are working to the same brief. Independent raters, either with experience in the area or with sufficient training from the aforementioned group of experts, would then manually code a selection of therapy transcripts for the selected features. Comparisons could then be made between the results of manual coding and computerised coding to assess overall accuracy rates as well as sensitivity and specificity scores to establish if the query is too broad or narrow.

## 11.5 Clinical implications

The exploratory nature of this project means that it is difficult to make solid clinical claims. Though predictive models of outcome and multiple linguistic features were statistically significant in development, this does not mean they are clinically significant. In models in which the linguistic features explain under 4% additional variation in the outcome data, it seems that the clinical significance of these features is very limited and the results of the validation cast even further doubt on their value. Furthermore, the range of error in predicting outcome scores is at present too large and would be too much of a risk if the developed model were applied as a clinical tool.

However, a number of the results indicate potential clinical applications. The linguistic features that were recurrent in models of mental health outcome suggest that there is definitely some value in measuring affect expressed in language. Though further refining on the measurement method, in particular in the case of CTS-R based features, and further adjusting of models would be recommended, this type of prediction work could be run in the background of ongoing treatment, without affecting or changing clinical practice, so as to monitor how the models perform in practice and whether they adapt within different populations. Finally, the finding that positive language from the therapist appears to act as a protective factor against drop out has the potential to affect clinical practice. Awareness of this association, and the role of therapist affect overall through the sharing of this information with therapists in training may improve practice. Furthermore, both consulting therapists about the link between positive language use and drop out, as well as undertaking qualitative analysis of transcripts, could provide further information about the direction of association between drop out and therapist positive language. This in turn might suggest how this mechanism could be used to improve therapy adherence.

## 11.6 Strengths and limitations

The results of this work need to be interpreted with awareness of its strengths and limitations. There are four major areas that characterise this work and each of these brings its strengths and challenges.

### 11.6.1 Originality of the project and its exploratory nature

One strength of this piece of work resides in its originality. It is a new approach to mental health research within a rapidly developing field with a great deal of potential. This means that the work described in this thesis has covered new ground in applying text mining methods to transcripts from online cognitive behaviour therapy, but also opened up a range of possibilities for future work (see section 11.7) with important clinical implications. However, the exploratory nature also posed a number of challenges in developing the methods to apply. There was very little precedent for this kind of research (if any in the case of text mining) and in particular in work that has analysed textual data from direct therapeutic exchanges. The lack of previous research to rely on meant that there was less certainty over a number of elements of the methods to be applied, sometimes leading to difficult decision-making. The most difficult of these decisions was in the selection of features for query development, such as whether to pursue the study of affective language beyond LIWC categories and which features of the CTS-R to focus on (see 11.1.2 for further details). This was made difficult by the lack of precedent in text mining research. Previous work that had been carried out using the I2E system tended to focus on narrow content questions such as ‘Do these X-ray report notes suggest the presence of pneumonia or not?’ and feature selection was handled by the development of a taxonomy of relevant terms by a team of health professionals (Liu et al., 2013). Additionally, there was no precedent of using the system within patient natural language. This created uncertainty around how to proceed and thus made for a difficult decision as an approach to follow needed to be selected. Nonetheless, this exploration of how text mining methods can be applied should assist in future work by setting a new precedent.

### 11.6.2 Data

A second characteristic of this project was the source of the data set used for analysis. The set of transcripts from online cognitive behaviour therapy, and associated demographics and outcome scores recorded for every session, is valuable for a number of reasons.  Its origins in clinical practice allow insight into how therapy is currently being provided and results that are directly interpretable within the context of the service. This means that work stemming from the analysis of these data could, and most likely will, have an impact on the future service provided by Ieso Digital Health and other providers of cognitive behaviour therapy either on or offline. The nature of the data set and the snapshot it provides of mental health services stand in contrast to work that aims to draw conclusions about language in individuals with a mental health condition from social media data, or about therapeutic characteristics from a much used repository of transcripts from psychological therapy sessions that includes a number of potentially outdated samples (e.g. transcripts from sessions with Carl Rogers or Albert Ellis). Much of the social media data used, for example, are language data that are public and individuals are often labeled with a mental health condition, or age and gender information, based on manual reading of what they post as opposed to any clinical diagnosis or collected data. The data set made available by Ieso for this project was therefore a very valuable one, and the source of much the envy for many researchers within computational linguistics. The data set also allowed the analyses to be carried out on natural language rather than language generated within experimental circumstances.

However, there are a number of limitations to this data set, most of which are directly linked to the origin of the data set as clinical data on which secondary analyses are being performed. As the data were collected as part of routine practice, and prior to the beginning of the project in the case of the development data set, the specific measures recorded were out of my control. Age was, for example, recorded as a categorical measure and there was no reporting of ethnicity, nationality, or first language, pieces of information that could have a large influence on the type of language an individual uses and the relationship between language use and mental health outcomes. Furthermore, the diagnosis provided with the dataset was primarily either a provisional diagnosis provided by the patient's GP as opposed to a formal mental health diagnosis or one provided in triage through a telephone assessment. This leaves room for error, especially where multiple conditions are present. Finally, it became apparent that PHQ-9 and GAD-7 scores were not always above the threshold for 'caseness' (i.e. the point at which therapy would normally be recommended) at the time of their first appointment. This was particularly the case in the development set and may be associated with the fact that many individuals in this population were on a waiting list for a long time.

A final potential limitation associated with the data is that little pre-processing was carried out prior to analysis. The transcripts were extracted directly from service records, with no corrections or editing carried out aside from the anonymisation process. This means that any spelling errors, 'text speak' or grammatical inconsistencies were kept within the data. Though I2E has some capacity to deal with spelling errors, it is possible that errors or alternate spellings used may have affected the measurement of linguistic features. The use of 'text speak' or uncommon contractions by therapists is actually something that was addressed by the service provider between the time of collection of the development and validation sets, meaning that the language in the validation set is generally cleaner and clearer than in the development set. There are a number of approaches to pre-processing data of this type, some of which involve removing the most common words, known as stopwords, but in this case the data were left in their original form as much as possible as the long-term goal for the service provider is to develop an automatic tool for use within the service. Keeping the textual data in their original format therefore was both an advantage and disadvantage in this project. The results are more representative of model performance in natural language but the variable quality of the written text due to misspellings or uncommon contractions is likely to have increased the noise in the data, leading to less clear results.

A final strength of this project in relation to data is the availability of a validation data set from a different year and geographical population than the development set, with which statistical models could be tested. This will be further covered in the section on strengths and weaknesses of the statistical approach. Working with two separate data sets also provided a second perspective from which to consider and look back on the approach taken throughout the research project. The lack of strength of the linguistic features in the external validation of the model raised questions about whether the queries developed had performed well in the second data set. These were developed closely with the development data set and when features did not validate well in the second data set, the possibility arose that these were too specific to the first dataset. Despite accessing the same service, differences in geographical location, service developments and even differences in waiting times, may have impacted the data in a greater way than anticipated. In future, it may be worth considering using a range of data from different local services within Ieso. Though this was more difficult at the time as the service was only accessible in a handful of areas, this has been rolled out across a large number of locations and would be more feasible now.

### 11.6.3 Text mining approach

The use of a text mining approach brought its own set of strengths and weaknesses to the project. Its subjective nature and the manual method of development meant that the subsequently developed queries were highly interpretable and have comprehensible meaning. This is in contrast with a number of machine learning methods where the linguistic analysis is much more of a black box with sets of predictors being put forward that are often difficult to interpret. In some cases, clusters of words are presented as topics, and a general theme for these may be determined but individual terms are often only loosely connected with this. The approach used here involved the development of specific text mining queries around a given theme or aspect of language or therapy. The association of these with outcomes was then evaluated statistically leading to results that are directly interpretable.

However, the text mining approach comes with its own challenges, primarily associated with the manual and therefore subjective development of the features selected for analysis. As the method requires the development of specific language features, a choice needed to be made prior to analysis and development of these features of what to focus on. In the case of this project, there was little previous work on which to guide this choice, and none within text mining and therapy transcripts. Previous work on language and mental health has focused primarily on the application of the LIWC dictionary, so this was deemed to be the best starting point. Additionally, the development of a query and its success in accurately measuring the feature in question relies on personal expertise and understanding of the data and linguistic features being studied.

Within previous text mining work, there are a number of different approaches that researchers use in order to reduce the impact of subjective query development. In terms of the development of a dictionary of terms from which to work, one option is to use an annotated data set from which terms are extracted, either manually with the assistance of experts in the field (Liu et al., 2013) or through machine learning methods, considered objective, such as topic modelling (Imel et al., 2015). Alternately, in an unannotated text, ‘named entity recognition’ programmes can be run over text in order to extract common ‘entities’ (terms) from which to form a dictionary (Zhu et al., 2013). However, these tend to depend on the nature of the task and data at hand and are not applicable in all situations. In drug discovery tasks, for example, where the goal is to determine novel associations between two entities, statistical co-occurence may be an appropriate and objective approach. When working with natural language and concepts such as affect, such objectivity is difficult. In this case, using separate datasets for development and testing of queries is a common method to validate queries as well as compare query performance to a gold standard, often a manual annotation. Such separate sensitivity analyses of individual features were not carried out during the project as each query developed was conducted through an iterative process of editing the query and checking the obtained results. However, this does limit the validity of the results and it would be beneficial to carry out some form of sensitivity analyses on the features that appear significant in the developed models with the assistance of individuals not involved in the query development process.

Though there is little comparable work to be found in the literature, some of these challenges appear to be echoed in research into the applications of text mining more broadly in biomedical research. In a paper looking at the applications of text mining in cancer research, Zhu et al., (2013) highlight that throughout applications of text mining in biomedical research, the development of a comprehensive set of keywords or terms with which to begin query development appears to be a consistent challenge with many research groups looking to combine expert knowledge, an annotated source of terms and machine learning methods to harvest new terms from the literature (Zhu et al., 2013). The research work that does apply text mining methods to mental health data appear to focus on one area of feature identification. Yu et al. (2011) for example, focus on identifying negative life events within the categories of family, work, love, school and social in online posts (Yu et al., 2011), while Wu et al. (2012) focus on the extraction of word pairs that indicate a causal relationship from a similar data set of online self narratives (Wu, Yu, & Chang, 2012). These pieces of work can be seen to support the idea of an analytical approach split into discrete stages as described here.

### 11.6.4 Statistical analyses

The final important element of this project was the statistical analysis. Given the continuous outcome data, the development of risk prediction models led to more interpretable models of outcome scores than the development of classification tasks as is more common in the computational linguistics literature. Including random effects in the models allowed for clustering in the data to be taken into account and therefore to develop more accurate models. The detailed analysis permitted by risk prediction models is a strength of this project. There are, however, some elements of the statistical analysis that can be seen as limitations of the project.

Primarily, the number of linguistic features considered and tested within the project and within the same data set was very large. The work carried out was intended to be exploratory and features were tested in sets so as to allow enough data within individual models but the total number of features tested to predict the same outcome was nonetheless large. This makes a false positive result, or the suggestion of a significant association where there isn't one, more likely and puts forward the necessity of replication of these results before drawing firm conclusions. The contribution of the linguistic features to the models developed was generally weak; it therefore seems unnecessary to look to quantify the probability of a false positive result at this stage. Caution is also applied in interpreting these results. Furthermore, the primary aims of this thesis were around the development of prediction models, assessed based on R-squared and calibration slope values.

External validation of the models developed using a distinct data set was the primary strategy applied to remedy this concern. Additionally, models were recalibrated to the new data set. This step provided evidence of any differences in the associations between linguistic features and outcome scores between the two data sets. A number of linguistic features that had been significantly associated with outcome in models fitted on the development set were not so in the validation set suggesting that these associations were not generalisable to a different population. Nonetheless, where re-calibration supported the significant association of variables with outcome in both data sets, this strengthens the evidence of those associations. This is particularly the case for affect-based features, for example. The testing of multiple feature sets does strengthen conclusions when the association between a language features and mental health outcomes is significant across different measurement methods. For example, patient positive language in models of PHQ-9 and GAD-7 scores at the end of treatment when measured using LIWC, LIWC-based queries and PANAS-X based affective measures (4.3.6, 4.3.7, 5.2.6, 5.2.7, 6.2.6, 6.2.7).

## 11.7 Future directions

Over the course of this project, through the analysis and interpretation of results and in the understanding of other work in the field that has recently or is in the process of being carried out, some clear ideas of potential future directions for research in this area became apparent. When applying text mining methods to textual clinical data, it appears that focusing on identifying and extracting specific pieces of information may be the best approach to take. Specific here refers to smaller, less ambiguous units of information than might be the case for affect, for example. One possible approach that may be well-suited to text mining methods is that of considering transcripts from online cognitive behaviour therapy, as mines of information to be extracted in the same way that electronic health records have been approached in recent years. This would mean extracting information from the text to answer questions such as 'did the therapist set an agenda?' Or 'does this patient mention a history of mental illness in their family?' These extracted measures could then be used within predictive models as was done throughout this project. Taking this approach would require a large amount of work in its set up but would also create potential for a whole new set of research.

Before undertaking any particular data extraction, however, agreement would be required on which elements of information in a session transcript might be both useful and recognizable in terms of language patterns. Qualitative work such as that carried out by Ekberg et al. (2015) should be involved here in order both to understand what information is contained within the transcripts, but also what types of information an expert considers might be useful in understanding or predicting therapy outcome. This might range from further development of CTS-R type items, and whether they are present in treatment, to extraction of information about a patient's family history or medication. One example of an area that has been discussed for future work with Ieso Digital Health is identifying whether therapists are applying specific protocol methods when treating patients. Therapists are trained to select a specific protocol for treatment and to keep to this as opposed to jumping between different protocol methods with the same patient. A Beckian protocol approach to treating depression, for example, relies on a set of techniques based within Beck's theory of cognitive therapy for depression. If the presence of specific techniques or change mechanisms in session transcripts can be reliably detected, further work could both be carried out to understand whether the consistency of these affects patient outcome or even if specific mechanisms are more useful than others in this format. This is just one example of the type of data that could potentially be extracted, but it is important to remember that each feature to be extracted would need careful development.

The end product of this work could be a whole new set of features extracted from the data with which further research and analysis could be carried out. In the same way that a patient may be given multiple questionnaires or scales to complete during a research project, transcripts from treatment could be read by a series of linguistic tools that aim to identify different aspects of their history, treatment and therapy so as to extract information and enter it into a database without requiring the patient to report it again within a questionnaire. The potential within research of this type of database is clear but it could also have important clinical implications. A database alone could improve care provision by providing more adapted or personalized care but also increase continuity of care if there is a change in service provider, or at the end of treatment. Furthermore, the research carried out using this approach would most likely improve understanding of what works for whom within this form of cognitive behaviour therapy and therefore improve tailoring of service provision to each patient.

Ieso Digital Health are providing text based online cognitive behaviour therapy which relies on an instant messaging platform to deliver the treatment. This means that collection of transcripts from treatment is routine. The resources are therefore there and the tools with which to best exploit them need to be further developed. As long as patients following treatment are comfortable with their anonymised data being used for research in this way, it seems that there are an enormous number of possibilities for further research applying text mining methods within this data format. The results of this further research have the potential to lead to more individualised care and greater allocation of resources to those who require them or would benefit from them most in this context.

## Conclusions

This research brought together psychological therapy and linguistic analysis in an original project that looked to explore the potential of applying text mining methods to the analysis of transcripts from online cognitive behaviour therapy. Four sets of linguistic features were developed emerging from a variety of sources; a dictionary that has been previously used in mental health research, a dictionary developed specifically for this project and an evaluative scale of therapist skill and adherence to the CBT structure. The associations between these linguistic features and mental health outcome measures were explored separately and predictive models of therapy outcomes were then developed. These were then tested on an independent dataset. At this stage and despite previous significant associations between linguistic features and outcome scores, the value of including linguistic features in the models fell to being negligible or non-existent. The features investigated in this thesis therefore did not come across as strong markers of mental health state or strong predictors of mental health outcomes.

Nonetheless, the work carried out throughout this research has led to a greater understanding of the potential applications of text mining within this data format and the processes that should be followed. A focus on narrower features may be well suited to this form of analysis as long as features are clearly defined. Involvement of a multidisciplinary team made up of linguistics, mental health and text mining experts would be beneficial, as would a formal process to test how well each feature is being measured. Features developed thus could also be considered as potential markers of mental health outcomes and be researched as such. The analyses carried out in this research project cover only a sample of many possible features to explore within this data set, the primary challenge for the future is to define clearly which to focus on next.

## References


Ahmad, N. Y., & Farrell, M. H. (2014). Linguistic markers of emotion in mothers of sickle cell carrier infants: What are they and what do they mean? *Patient Education and Counseling*, *94*(1), 128–133. https://doi.org/10.1016/j.pec.2013.09.021

Alparone, F., Caso, S., Agosti, A., & Rellini, A. (2004). *The Italian LIWC2001 Dictionary.* Austin: TX: LIWC.net.

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Barr Taylor, C. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, *21*(2), 361–376. https://doi.org/10.1016/j.chb.2004.02.008

Alvarez-Conrad, J., Zoellner, L. A., & Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, *15*(7), S159–S170. https://doi.org/10.1002/acp.839

Ambler, G., Seaman, S., & Omar, R. Z. (2012). An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine*, *31*(11–12), 1150–1161. https://doi.org/10.1002/sim.4371

Anderson, B., Goldin, P. R., Kurita, K., & Gross, J. J. (2008). Self-representation in social anxiety disorder: Linguistic analysis of autobiographical narratives. *Behaviour Research and Therapy*, *46*(10), 1119–1125. https://doi.org/10.1016/j.brat.2008.07.001

Anderson, T., Bein, E., Pinnell, B., & Strupp, H. (1999). Linguistic Analysis of Affective Speech in Psychotherapy: A case grammar approach. *Psychotherapy Research*, *9*(1), 88–99. https://doi.org/10.1080/10503309912331332611

Anderson, T., Ogles, B. M., Patterson, C. L., Lambert, M. J., & Vermeersch, D. A. (2009). Therapist effects: facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology*, *65*(7), 755–768. https://doi.org/10.1002/jclp.20583

Andersson, G., & Cuijpers, P. (2009). Internet-Based and Other Computerized Psychological Treatments for Adult Depression: A Meta-Analysis. *Cognitive Behaviour Therapy*, *38*(4), 196–205. https://doi.org/10.1080/16506070903318960

Arntz, A., Hawke, L. D., Bamelis, L., Spinhoven, P., & Molendijk, M. L. (2012). Changes in natural language use as an indicator of psychotherapeutic change in personality disorders. *Behaviour Research and Therapy*, *50*(3), 191–202. https://doi.org/10.1016/j.brat.2011.12.007

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49.

Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*, *12*, 82. https://doi.org/10.1186/1471-2288-12-82

Bados, A., Balaguer, G., & Saldaña, C. (2007). The efficacy of cognitive–behavioral therapy and the problem of drop-out. *Journal of Clinical Psychology*, *63*(6), 585–592. https://doi.org/10.1002/jclp.20368

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, *75*(6), 842–852. https://doi.org/10.1037/0022-006X.75.6.842

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*, *21*(1), 79–88. https://doi.org/10.1037/a0014643

Beck, A. T. (1979). *Cognitive Therapy of Depression*. Guilford Press.

Beck, J. S. (2010). Cognitive Therapy. In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470479216.corpsy0198/abstract

Beltman, M. W., Voshaar, R. C. O., & Speckens, A. E. (2010). Cognitive-behavioural therapy for depression in people with a somatic disease: meta-analysis of randomised controlled trials. *The British Journal of Psychiatry*, *197*(1), 11–19. https://doi.org/10.1192/bjp.bp.109.064675

Bergmann, B., Villmann, T., & Gumz, A. (2008). Vom Chaos zur Einsicht – Die Charakterisierung der Dynamik therapeutischer Veränderungsprozesse mittels textanalytischer Untersuchung von Verbatimprotokollen. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, *58*(09/10), 379–386. https://doi.org/10.1055/s-2007-986360

Berth, H. (2001). [The measurement of anxiety through computerized content analysis--automation of the Gottschalk-Gleser Test]. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *51*(1), 10–16.

Biber, D. (2009). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (Vol. 1).

Bjekic, J., Lazarevic, L., Zivanovic, M., & Knezevic, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis - LIWCser. *Psihologija*, *47*(1), 5–32. https://doi.org/10.2298/PSI1401005B

Blackburn, I.-M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The Revised Cognitive Therapy Scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy*, *29*(4), 431–446. https://doi.org/10.1017/S1352465801004040

Brockmeyer, T., Holtforth, M. G., Bents, H., Kämmerer, A., Herzog, W., & Friederich, H.-C. (2012). Starvation and emotion regulation in anorexia nervosa. *Comprehensive Psychiatry*, *53*(5), 496–501. https://doi.org/10.1016/j.comppsych.2011.09.003

Brown, L. A., Craske, M. G., Glenn, D. E., Stein, M. B., Sullivan, G., Sherbourne, C., … Rose, R. D. (2013). CBT competence in novice therapists improves anxiety outcomes. *Depression and Anxiety*, *30*(2), 97–115. https://doi.org/10.1002/da.22027

Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2012). A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using

Linguistic Features. In *Interspeech* (pp. 2254–2257). Retrieved from http://www-scf.usc.edu/~dogancan/files/dogancan-is12.pdf

Carlbring, P., Nilsson-Ihrfelt, E., Waara, J., Kollenstam, C., Buhrman, M., Kaldo, V., … Andersson, G. (2005). Treatment of panic disorder: live therapy vs. self-help via the Internet. *Behaviour Research and Therapy*, *43*(10), 1321–1333. https://doi.org/10.1016/j.brat.2004.10.002

Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., … Belliveau, R. A. (2015). Validation of Electronic Health Record Phenotyping of Bipolar Disorder Cases and Controls. *American Journal of Psychiatry*, *172*(4), 363–372. https://doi.org/10.1176/appi.ajp.2014.14030423

Clore, G. L., Ortony, A., & Foss, M. A. (1987). The Psychological Foundations of the affect lexicon. *Journal of Personality and Social Psychology*, *53*(4), 751–766.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. https://doi.org/10.1111/j.0956-7976.2004.00741.x

Colby, K. M., Weber, S., & Hilf, F. D. (1971). Artificial Paranoia. Artificial Intelligence.

Collins, G. S., Ogundimu, E. O., & Altman, D. G. (2016). Sample size considerations for the external validation of a multivariable prognostic model: a resampling study: Sample size considerations for validating a prognostic model. *Statistics in Medicine*, *35*(2), 214–226. https://doi.org/10.1002/sim.6787

Consedine, N. S., Krivoshekova, Y. S., & Magai, C. (2012). Play It (Again) Sam: Linguistic Changes Predict Improved Mental and Physical Health Among Older Adults. *Journal of Language and Social Psychology*, *31*(3), 240–262. https://doi.org/10.1177/0261927X12446736

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. *NAACL HLT 2015*, 1.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. *NAACL HLT 2015*, 31.

Corrigan, P. W., & Rao, D. (2012). On the Self-Stigma of Mental Illness: Stages, Disclosure, and Strategies for Change. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *57*(8), 464–469.

Cuijpers, P., Marks, I. M., van Straten, A., Cavanagh, K., Gega, L., & Andersson, G. (2009). Computer-aided psychotherapy for anxiety disorders: a meta-analytic review. *Cognitive Behaviour Therapy*, *38*(2), 66–82. https://doi.org/10.1080/16506070802694776

Cuijpers, P., van Straten, A., Schuurmans, J., van Oppen, P., Hollon, S. D., & Andersson, G. (2010). Psychotherapy for chronic major depression and dysthymia: A meta-analysis. *Clinical Psychology Review*, *30*(1), 51–62. https://doi.org/10.1016/j.cpr.2009.09.003

D'Andrea, W., Chiu, P. H., Casas, B. R., & Deldin, P. (2012). Linguistic Predictors of Post-Traumatic Stress Disorder Symptoms Following 11 September 2001. *Applied Cognitive Psychology*, *26*(2), 316–323. https://doi.org/10.1002/acp.1830

Dao, B., Nguyen, T., Phung, D., & Venkatesh, S. (2014). Effect of Mood, Social Connectivity and Age in Online Depression Community via Topic and

Linguistic Analysis. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, & Y. Zhang (Eds.), *Web Information Systems Engineering – WISE 2014* (pp. 398–407). Springer International Publishing. https://doi.org/10.1007/978-3-319-11749-2_30

Department of Health. (2011). IAPT outline service specification. Retrieved 11 April 2016, from http://www.iapt.nhs.uk/silo/files/iapt-outline-service-specification.pdf

Department of Health. (2012, November). IAPT three-year report: The first million patients. Retrieved 11 April 2016, from http://www.iapt.nhs.uk/silo/files/iapt-3-year-report.pdf

Di Giulio, G. (2010). *Therapist, client factors, and efficacy in cognitive behavioural therapy a meta-analytic exploration of factors that contribute to positive outcome.* Library and Archives Canada = Bibliothèque et Archives Canada, Ottawa.

Ekberg, S., Barnes, R. K., Kessler, D. S., Mirza, S., Montgomery, A. A., Malpass, A., & Shaw, A. R. G. (2015). Relationship between Expectation Management and Client Retention in Online Cognitive Behavioural Therapy. *Behavioural and Cognitive Psychotherapy*, *43*(6), 732–743. https://doi.org/10.1017/S1352465814000241

Federoff, I. C., & Taylor, S. (2007). Psychological and Pharmacological Treatments of Social Phobi... : Journal of Clinical Psychopharmacology. Retrieved 11 April 2016, from http://journals.lww.com/psychopharmacology/Fulltext/2001/06000/Psychological_and_Pharmacological_Treatments_of.11.aspx

Fertuck, E. A., Bucci, W., Blatt, S. J., & Ford, R. Q. (2004). Verbal Representation and Therapeutic Change in Anaclitic and Introjective Inpatients. *Psychotherapy: Theory, Research, Practice, Training*, *41*(1), 13–25. https://doi.org/10.1037/0033-3204.41.1.13

Fertuck, E. A., Mergenthaler, E., Target, M., Levy, K. N., & Clarkin, J. F. (2012). Development and criterion validity of a computerized text analysis measure of reflective functioning. *Psychotherapy Research*, *22*(3), 298–305. https://doi.org/10.1080/10503307.2011.650654

Fontao, M. I., & Mergenthaler, E. (2008). Therapeutic factors and language patterns in group therapy application of computer-assisted text analysis to the examination of microprocesses in group therapy: Preliminary findings. *Psychotherapy Research*, *18*(3), 345–354. https://doi.org/10.1080/10503300701576352

Fredrickson, B. L., & Joiner, T. (2002). Positive Emotions Trigger Upward Spirals Toward Emotional Well-Being. *Psychological Science*, *13*(2), 172–175. https://doi.org/10.1111/1467-9280.00431

Galor, S., & Hentschel, U. (2009). Analysis of suicidal behaviour in Israeli veterans and terror victims with post-traumatic stress disorder by using the computerised Gottschalk-Gleser scales. *Clinical Psychologist*, *13*(3), 102–110. https://doi.org/10.1080/13284200903353072

Gamber, A. M., Lane-Loney, S., & Levine, M. P. (2013). Effects and Linguistic Analysis of Written Traumatic Emotional Disclosure in an Eating-Disordered Population. *The Permanente Journal*, *17*(1), 16–20. https://doi.org/10.7812/TPP/12-056

Garfield, D. A. S., Rapp, C. R., & Evens, M. (1992). Natural Language Processing in Psychiatry. Artifical Intelligence Technology and Psychopathology. *The Journal of Nervous and Mental Disease*, *180*(4), 227–237.

Gil, P. J. M., Xavier, F., & Meca, J. S. (2001). Effectiveness of cognitive-behavioural treatment in social phobia: A meta-analytic review. *Psychology in Spain*, *5*(1), 17–25.

Gottschalk, L. A., & Bechtel, R. J. (1982). The measurement of anxiety through the computer analysis of verbal samples. *Comprehensive Psychiatry*, *23*(4), 364–369. https://doi.org/10.1016/0010-440X(82)90086-4

Gottschalk, L. A., & Hoigaard-Martin, J. (1985). A depression scale applicable to verbal samples. *Psychiatry Research*, *17*, 213–227.

Gottschalk, L. A., Stein, M. K., & Shapiro, D. H. (1997). The application of computerized content analysis of speech to the diagnostic process in a psychiatric outpatient clinic. *Journal of Clinical Psychology*, *53*(5). Retrieved from http://escholarship.org/uc/item/35m89370

Green, S. A., Honeybourne, E., Chalkley, S. R., Poots, A. J., Woodcock, T., Price, G., … Green, J. (2015). A retrospective observational analysis to identify patient and treatment-related predictors of outcomes in a community mental health programme. *BMJ Open*, *5*(5), e006103. https://doi.org/10.1136/bmjopen-2014-006103

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Harrell, F., E. ..Jr. (2001). Multivariable Modeling Strategies. In *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression and Survival Analysis.* New York: Springer.

Haug, S., Strauss, B., Gallas, C., & Kordy, H. (2008). New prospects for process research in group therapy: Text-based process variables in psychotherapeutic Internet chat groups. *Psychotherapy Research*, *18*(1), 88–96. https://doi.org/10.1080/10503300701368008

Hayeri, N., Chung, C. K., Booth, R. J., & Pennebaker, J. W. (2010). *LIWC for Arabic texts*. Austin: TX: LIWC.net.

He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*(3), 441–447. https://doi.org/10.1016/j.psychres.2012.01.032

HM G. (2011). No health without mental health: a cross-government mental health outcomes strategy for people of all ages. London: Department of Health. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213761/dh_124058.pdf

Holmes, D., Alpers, G. W., Ismailji, T., Classen, C., Wales, T., Cheasty, V., … Koopman, C. (2007). Cognitive and Emotional Processing in Narratives of Women Abused by Intimate Partners. *Violence Against Women*, *13*(11), 1192–1205. https://doi.org/10.1177/1077801207307801

Howes, C., Purver, M., & McCabe, R. (2013). Using Conversation Topics for Predicting Therapy Outcomes in Schizophrenia. *Biomedical Informatics Insights*, *6*(Suppl 1), 39–50. https://doi.org/10.4137/BII.S11661

Howes, C., Purver, M., & McCabe, R. (2014). Linguistic indicators of severity and progress in online text-based therapy for depression. *ACL 2014*, 7.

Howes, C., Purver, M., McCabe, R., Healey, P. G., & Lavelle, M. (2012a). Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue* (p. 155). Retrieved from http://hal-sfo.ccsd.cnrs.fr/hal-01138035/document#page=164

Howes, C., Purver, M., McCabe, R., Healey, P. G., & Lavelle, M. (2012b). Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 79–83). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2392814

Huppert, J. D., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2006). The Interaction of Motivation and Therapist Adherence Predicts Outcome in Cognitive Behavioral Therapy for Panic Disorder: Preliminary Findings. *Cognitive and Behavioral Practice*, *13*(3), 198–204. https://doi.org/10.1016/j.cbpra.2005.10.001

Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, *23*(2), 8–28.

IAPT Data Handbook. (2011, June). Retrieved 20 May 2016, from http://www.iapt.nhs.uk/silo/files/iapt-data-handbook-v2.pdf

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, *52*(1), 19–30. https://doi.org/10.1037/a0036841

Junghaenel, D. U., Smyth, J. M., & Santner, L. (2008). Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology*, *27*(1), 36–55.

Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, *130*(6), 515–524.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263–286.

Kailer, A., & Chung, C. K. (2010). *The Russian LIWC2007 dictionary*. Austin: TX: LIWC.net.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276. https://doi.org/10.1093/schbul/13.2.261

Keen, A. J. A., & Freeston, M. H. (2008). Assessing competence in cognitive–behavioural therapy. *The British Journal of Psychiatry*, *193*(1), 60–64. https://doi.org/10.1192/bjp.bp.107.038588

Kessler, D., Lewis, G., Kaur, S., Wiles, N., King, M., Weich, S., … Peters, T. J. (2009). Therapist-delivered internet psychotherapy for depression in primary care: a randomised controlled trial. *The Lancet*, *374*(9690), 628–634. https://doi.org/10.1016/S0140-6736(09)61257-5

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of*

*General Psychiatry*, *62*(6), 593–602. https://doi.org/10.1001/archpsyc.62.6.593

Kessler, R. C., de Jonge, P., Shahly, V., van Loo, H. M., Wang, P. S. E., Wilcox, M. A., … Hammen, C. L. (2010). *Handbook of depression* (2nd ed.). Guilford Press.

Kessler, R. C., Ruscio, A. M., Shear, K., & Wittchen, H.-U. (2009). *Epidemiology of Anxiety Disorders*. (M. M. Antony & M. B. Stein, Eds.). Oxford University Press. Retrieved from //www.oxfordhandbooks.com/10.1093/oxfordhb/9780195307030.001.0001/oxfordhb-9780195307030-e-2

Kiropoulos, L. A., Klein, B., Austin, D. W., Gilson, K., Pier, C., Mitchell, J., & Ciechomski, L. (2008). Is internet-based CBT for panic disorder and agoraphobia as effective as face-to-face CBT? *Journal of Anxiety Disorders*, *22*(8), 1273–1284. https://doi.org/10.1016/j.janxdis.2008.01.008

Konovalov, S., Scotch, M., Post, L., & Brandt, C. (2010). Biomedical Informatics Techniques for Processing and Analyzing Web Blogs of Military Service Members. *Journal of Medical Internet Research*, *12*(4), e45. https://doi.org/10.2196/jmir.1538

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 357–361. https://doi.org/10.1037/0033-3204.38.4.357

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*(2), 165–174. https://doi.org/10.1002/jclp.20113

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The Effects of Providing Therapists With Feedback on Patient Progress During Psychotherapy: Are Outcomes Enhanced? *Psychotherapy Research*, *11*(1), 49–68. https://doi.org/10.1080/713663852

Lee, H. S., & Cohn, L. D. (2009). Assessing coping strategies by analysing expressive writing samples. *Stress and Health*, *26*(3), 250–260. https://doi.org/10.1002/smi.1293

Lewinsohn, P. M., Rohde, P., Seeley, J. R., & Fischer, S. A. (1991). Age and depression: Unique and shared effects. *Psychology and Aging*, *6*(2), 247–260. https://doi.org/10.1037/0882-7974.6.2.247

Liehr, P., Marcus, M. T., Carroll, D., Granmayeh, L. K., Cron, S. G., & Pennebaker, J. W. (2010). Linguistic Analysis to Assess the Effect of a Mindfulness Intervention on Self-Change for Adults in Substance Use Recovery. *Substance Abuse*, *31*(2), 79–85. https://doi.org/10.1080/08897071003641271

Liu, V., Clark, M. P., Mendoza, M., Saket, R., Gardner, M. N., Turk, B. J., & Escobar, G. J. (2013). Automated identification of pneumonia in chest radiograph reports in critically ill patients. *BMC Medical Informatics and Decision Making*, *13*(1), 1.

Lo Verde, R., Sarracino, D., & Vigorelli, M. (2012). Therapeutic Cycles and Referential Activity in the Analysis of the Therapeutic Process. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *15*(1), 22. https://doi.org/10.4081/ripppo.2012.99

Lorenz, T. A., & Meston, C. M. (2012). Associations among childhood sexual abuse, language use, and adult sexual functioning and satisfaction. *Child Abuse & Neglect*, *36*(2), 190–199. https://doi.org/10.1016/j.chiabu.2011.09.014

Luborsky, L., Mintz, J., Auerbach, A., Christoph, P., Bachrach, H., Todd, T., … O'Brien, C. P. (1980). Predicting the outcome of psychotherapy. findings of the Penn Psychotherapy Project. *Archives of General Psychiatry*, *37*(4), 471–481.

Lyons, E. J., Mehl, M. R., & Pennebaker, J. W. (2006). Pro-anorexics and recovering anorexics differ in their linguistic Internet self-presentation. *Journal of Psychosomatic Research*, *60*(3), 253–256. https://doi.org/10.1016/j.jpsychores.2005.07.017

Maher, B. A., McKean, K. O., & McLaughlin, B. (1966a). Studies in psychotic language. In P. J. Stone, D. C. Dunphy, & M. S. Smith (Eds.), *The General Inquirer: A computer approach to content analysis,*. Oxford, England: M.I.T. Press.

Maher, B. A., McKean, K. O., & McLaughlin, B. (1966b). Studies in psychotic language. In P. J. Stone (Ed.), *The General Enquirer: A Computer Approach to Content Analysis*. Boston, MA: MIT.

Martin, D. J., Garske, J. P., & Katherine, M. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*(3), 438–450. https://doi.org/10.1037/0022-006X.68.3.438

McCarthy, K. L., Mergenthaler, E., & Grenyer, B. F. S. (2014). Early in-session cognitive-emotional problem-solving predicts 12-month outcomes in depression with personality disorder. *Psychotherapy Research*, *24*(1), 103–115. https://doi.org/10.1080/10503307.2013.826834

McKenzie, D. P., Clarke, D. M., Forbes, A. B., & Sim, M. R. (2010). Pessimism, Worthlessness, Anhedonia, and Thoughts of Death Identify DSM–IV Major Depression in Hospitalized, Medically Ill Patients. *Psychosomatics*, *51*(4), 302–311. https://doi.org/10.1016/S0033-3182(10)70701-5

McNicoll, A. (2015, March 20). Mental health trust funding down 8% from 2010 despite coalition's drive for parity of esteem. Retrieved 11 April 2016, from http://www.communitycare.co.uk/2015/03/20/mental-health-trust-funding-8-since-2010-despite-coalitions-drive-parity-esteem/

Mergenthaler, E. (1996). Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, *64*(6), 1306–1315. https://doi.org/10.1037/0022-006X.64.6.1306

Mergenthaler, E. (1997). The therapeutic cycles model in psychotherapy research: Theory, measurement and clinical application. *European Psychiatry*, *12*, 143s. https://doi.org/10.1016/S0924-9338(97)80391-4

Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *British Journal of Medical Psychology*, *72*(3), 339–354.

Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. *NAACL HLT 2015*, 11.

Molendijk, M. L., Bamelis, L., van Emmerik, A. A. P., Arntz, A., Haringsma, R., & Spinhoven, P. (2010). Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behaviour Research and Therapy*, *48*(1), 44–51. https://doi.org/10.1016/j.brat.2009.09.007

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: a meta-analysis. *Psychological Bulletin*, *128*(4), 638–662.

Neuman, Y., Cohen, Y., Assaf, D., & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, *56*(1), 19–25. https://doi.org/10.1016/j.artmed.2012.06.001

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., & Berk, M. (2014). Affective and Content Analysis of Online Depression Communities. *IEEE Transactions on Affective Computing*, *5*(3), 217–226. https://doi.org/10.1109/TAFFC.2014.2315623

NICE. (2009). Depression in adults: recognition and management. Retrieved 11 April 2016, from https://www.nice.org.uk/guidance/cg90/chapter/1-guidance

NICE. (2011). Generalised anxiety disorder and panic disorder in adults: management. Retrieved 11 April 2016, from https://www.nice.org.uk/guidance/cg113/chapter/1-Guidance

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50.

Park, M., McDonald, D. W., & Cha, M. (2013). Perception Differences between the Depressed and Non-Depressed Users in Twitter. In *ICWSM*. Retrieved from http://pensivepuffin.com/dwmcphd/papers/Park.PerceptionDifferences.ICWSM2013.pdf

PCAD Manual. (2016). (Version 3) [Windows]. Retrieved from http://www.gb-software.com/PCADManual.pdf

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2001). *Linguistic Inquiry and Word Count (software and manual)*. Mahwah, NJ: Erlbaum Publishers. Retrieved from http://homepage.psy.utexas.edu/HomePage/Faculty/pennebaker/reprints/LIWC2001.pdf

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count. LIWC.net. Retrieved from http://homepage.psy.utexas.edu/HomePage/Faculty/pennebaker/reprints/LIWC2001.pdf

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*. Retrieved from https://utexas-ir.tdl.org/handle/2152/31333

Pennebaker, J. W., & King, L., A. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003a). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003b). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291–301. https://doi.org/10.1037/0022-3514.85.2.291

Pepinsky, H. B. (1978). A Computer-Assisted Language Analysis System (CALAS) and Its Applications. Retrieved from http://eric.ed.gov/?id=ED162663

Pepinsky, H. B. (1985). A metalanguage of text. In V. M. Rentel, S. A. Corson, & B. R. Dunn (Eds.), *Psychophysiological Aspects of Reading and Learning*. Taylor & Francis.

Perlis, R. H., Iosifescu, D. V., Castro, V. M., Murphy, S. N., Gainer, V. S., Minnier, J., … Smoller, J. W. (2012). Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological Medicine*, *42*(1), 41–50. https://doi.org/10.1017/S0033291711000997

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical Informatics Insights*, *2010*(3), 19–28.

Pestian, J., Pestian, J., Pawel Matykiewicz, Brett South, Ozlem Uzuner, & John Hurdle. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 3. https://doi.org/10.4137/BII.S9042

Pfeiffer, P. N., Heisler, M., Piette, J. D., Rogers, M. A. M., & Valenstein, M. (2011). Efficacy of peer support interventions for depression: a meta-analysis. *General Hospital Psychiatry*, *33*(1), 29–36. https://doi.org/10.1016/j.genhosppsych.2010.10.002

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Française*, *56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

Powers, M. B., Sigmarsson, S. R., & Emmelkamp, P. M. G. (2008). A Meta–Analytic Review of Psychological Treatments for Social Anxiety Disorder. *International Journal of Cognitive Therapy*, *1*(2), 94–113. https://doi.org/10.1680/ijct.2008.1.2.94

Priebe, S., & Gruyters, T. (1992). The Role of the Helping Alliance in Psychiatric Community Care. A prospective study. *The Journal of Nervous and Mental Disease*, *181*(9).

Probst, T., Lambert, M. J., Loew, T. H., Dahlbender, R. W., Göllner, R., & Tritt, K. (2013). Feedback on patient progress and clinical support tools for therapists: improved outcome for patients at risk of treatment failure in psychosomatic in-patient therapy under the conditions of routine practice. *Journal of Psychosomatic Research*, *75*(3), 255–261. https://doi.org/10.1016/j.jpsychores.2013.07.003

Pulverman, C. S., Lorenz, T. A., & Meston, C. M. (2015). Linguistic changes in expressive writing predict psychological outcomes in women with history of childhood sexual abuse and adult sexual dysfunction. *Psychological Trauma: Theory, Research, Practice, and Policy*, *7*(1), 50–57. https://doi.org/10.1037/a0036462

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482–491). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2380875

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In *ICWSM*. Retrieved from http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-020.pdf

Referential Activity (RA) - The Referential Process. (2015). Retrieved 29 April 2016, from http://www.thereferentialprocess.org/theory/referential-activity

Rellini, A. H., & Meston, C. M. (2007). Sexual Desire and Linguistic Analysis: A Comparison of Sexually-Abused and Non-Abused Women. *Archives of Sexual Behavior*, *36*(1), 67–77. https://doi.org/10.1007/s10508-006-9076-9

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. *NAACL HLT 2015*, 99.

Robbins, M. L., Mehl, M. R., Smith, H. L., & Weihs, K. L. (2013). Linguistic indicators of patient, couple, and family adjustment following breast cancer: Linguistic indicators of adjustment following breast cancer. *Psycho-Oncology*, *22*(7), 1501–1508. https://doi.org/10.1002/pon.3161

Rosenbach, C., & Renneberg, B. (2015). Remembering rejection: specificity and linguistic styles of autobiographical memories in borderline personality disorder and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 85–92. https://doi.org/10.1016/j.jbtep.2014.09.002

Rosenberg, S. D., & Tucker, G. J. (1976). Verbal Content and the diagnosis of Schizophrenia. Presented at the 29th Annual Meeting of the American Psychiatric Association, Florida.

Rosenheck, R., Leslie, D., Keefe, R., McEvoy, J., Swartz, M., Perkins, D., … Lieberman, J. (2006). Barriers to employment for people with schizophrenia. *American Journal of Psychiatry*, *163*(3), 411–417.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Ruhmland, M., & Margraf, J. (2001). Efficacy of psychological treatments for specific phobia and obsessive compulsive disorder. *Verhaltenstherapie*, *11*(1), 14–26. https://doi.org/10.1159/000050321

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Michael, T., Dobson, K. S., … Imber, S. D. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and*

*Clinical Psychology*, *67*(6), 837–846. https://doi.org/10.1037/0022-006X.67.6.837

Shickel, B., Heesacker, M., Benton, S., Ebadi, A., Nickerson, P., & Rashidi, P. (2016). Self-Reflective Sentiment Analysis. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 23–32). San Diego. Retrieved from http://hollingk.github.io/CLPsych/pdf/CLPsych3.pdf

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silva, M. J. D., McKenzie, K., Harpham, T., & Huttly, S. R. A. (2005). Social capital and mental illness: a systematic review. *Journal of Epidemiology and Community Health*, *59*(8), 619–627. https://doi.org/10.1136/jech.2004.029678

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Steine, S., Finset, A., & Laerum, E. (2001). A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome. *Family Practice*, *18*(4), 410–418. https://doi.org/10.1093/fampra/18.4.410

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Stillman, D., Kornblith, S., & Cheslack-Postava, F. (2013). Zotero (Version 4.0.29.11). Fairfax, VA, USA: Center for History and New Media, George Mason University.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484–498. https://doi.org/10.1002/bs.3830070412

Strupp, H. H. (1993). The Vanderbilt psychotherapy studies: synopsis. *Journal of Consulting and Clinical Psychology*, *61*(3), 431–433.

Strupp, H. H., & Hadley, S. W. (1979). Specific vs nonspecific factors in psychotherapy: A controlled study of outcome. *Archives of General Psychiatry*, *36*(10), 1125–1136. https://doi.org/10.1001/archpsyc.1979.01780100095009

Tanana, M., Hallgren, K., Imel, Z., Atkins, D., Smyth, P., & Srikumar, V. (2015). Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *NAACL HLT 2015*, 71.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

The King's Fund. (2015, February 25). Has the government put mental health on an equal footing with physical health? Retrieved 11 April 2016, from http://www.kingsfund.org.uk/projects/verdict/has-government-put-mental-health-equal-footing-physical-health

Thorndike, E. L., & Lorge, I. (1944). The Teacher's Word Book of 30,000 words. New Yok: Teachers College Press.

Thornicroft, G. (2011). Physical health disparities and mental illness: the scandal of premature mortality. *The British Journal of Psychiatry*, *199*(6), 441–442. https://doi.org/10.1192/bjp.bp.111.092718

Tolin, D. F. (2010). Is cognitive–behavioral therapy more effective than other therapies?A meta-analytic review. *Clinical Psychology Review*, *30*(6), 710–720. https://doi.org/10.1016/j.cpr.2010.05.003

Tolman, R. M., Himle, J., Bybee, D., Abelson, J. L., Hoffman, J., & Van Etten-Lee, M. (2009). Impact of Social Anxiety Disorder on Employment Among Women Receiving Welfare Benefits. *Psychiatric Services*, *60*(1), 61–66. https://doi.org/10.1176/ps.2009.60.1.61

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. *Psychological Assessment*, *25*(4), 1069–1078. https://doi.org/10.1037/a0033007

Trull, T. J., Useda, J. D., Conforti, K., & Doan, B.-T. (1997). Borderline Personality Disorder Features in Nonclinical Young Adults: 2. Two-Year Outcome. *Journal of Abnormal Psychology*, *106*(2), 307–314.

Tucker, G. J., & Rosenberg, S. D. (1975). Computer content analysis of schizophrenic speech: a preliminary report. *The American Journal of Psychiatry*, *132*(6), 611–616. https://doi.org/10.1176/ajp.132.6.611

Van der Zanden, R., Curie, K., Van Londen, M., Kramer, J., Steen, G., & Cuijpers, P. (2014). Web-based depression treatment: Associations of clients′ word use with adherence and outcome. *Journal of Affective Disorders*, *160*, 10–13. https://doi.org/10.1016/j.jad.2014.01.005

Vos T, Haby MM, Barendregt JJ, Kruijshaar M, Corry J, & Andrews G. (2004). THe burden of major depression avoidable by longer-term treatment strategies. *Archives of General Psychiatry*, *61*(11), 1097–1103. https://doi.org/10.1001/archpsyc.61.11.1097

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule-expanded form. Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Weiss, D. S., & Marmar, C. R. (1996). The Impact of Event Scale - Revised. In J. Wilson & T. M. Keane (Eds.), *Assessing psychological trauma and PTSD* (pp. 399–411). New York: Guilford.

Weissman, M., Leaf, P., Florio, L., Holzer, C., & Livingston, B. M. (1991). Affective disorders. In L. Robins & D. Regier (Eds.), *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study.* (pp. 53–80). New York: The Free Press;

Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Whisman, M. A. (1993). Mediators and moderators of change in cognitive therapy of depression. *Psychological Bulletin*, *114*(2), 248–265. https://doi.org/10.1037/0033-2909.114.2.248

Wiemer-Hastings, K., Janit, A. S., Wiemer-Hastings, P. M., Cromer, S., & Kinser, J. (2004). Automatic classification of dysfunctional thoughts: A feasibility test. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 203–212.

Wittchen, H. U., Zhao, S., Kessler, R. C., & Eaton, W. W. (1994). DSM-III-R generalized anxiety disorder in the National Comorbidity Survey. *Archives of General Psychiatry*, *51*(5), 355–364.

Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders*, *40*(8), 711–717. https://doi.org/10.1002/eat.20445

Wolf, M., Theis, F., & Kordy, H. (2013). Language Use in Eating Disorder Blogs Psychological Implications of Social Online Activity. *Journal of Language and Social Psychology*, *32*(2), 212–226. https://doi.org/10.1177/0261927X12474278

Wu, J.-L., Yu, L.-C., & Chang, P.-C. (2012). Detecting causality from online psychiatric texts using inter-sentential language patterns. *BMC Medical Informatics and Decision Making*, *12*(1), 72.

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). ‘ Rate My Therapist’: Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PloS One*, *10*(12), e0143055.

Young, J., & Beck, A. T. (1980). Cognitive therapy scale: Rating manual. *Unpublished Manuscript*, 36th.

Yu, L.-C., Chan, C.-L., Lin, C.-C., & Lin, I.-C. (2011). Mining association language patterns using a distributional semantic model for negative life event classification. *Journal of Biomedical Informatics*, *44*(4), 509–518. https://doi.org/10.1016/j.jbi.2011.01.006

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., … Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, *46*(2), 200–211. https://doi.org/10.1016/j.jbi.2012.10.007

Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie vand de ‘Linguistic Inquiry and Word Count’ (LIWC) [Dutch version of the Linguistic Inquiry and Word Count (LIWC]. *Gedrag & Gezonheid*, *32*(4), 271–281.

# Appendix A - Baseline model results

Prior to fitting models including linguistic features, a baseline model was fitted for each outcome score considered. Presented here are the developed baseline models for each of the outcome score that formed the base of the models presented in the main body of the thesis. These were developed following the same methods used in previous chapters, using backwards elimination and a significance threshold of p= 0.15 for inclusion in a model.

## A.1 Baseline outcome scores

Table A-1 presents the mean baseline PHQ-9 and GAD-7 scores in the development dataset. Within IAPT, the thresholds at which it is determined that an individual would benefit from psychological therapy are 10 on the PHQ-9 and 8 on the GAD-7. On average, the patients in this data set were above this threshold but there was a large range of baseline score spanning the full PHQ-9 and GAD-7 scales.

**Table A-1 Descriptive statistics for baseline outcome scores**

| Outcome score | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|
| **Baseline PHQ-9 score** | 12.60 | 6.46 | 0 | 27 |
| **Baseline GAD-7 score** | 12.00 | 5.29 | 0 | 21 |

## A.2 Mixed effects models results

### A.2.1 Outcome 1 – PHQ-9 score before session.

This model looks at the PHQ-9 score associated with a given appointment. The score at first appointment or assessment appointment attended by a patient was used as the baseline outcome score. The model includes data from 1906 appointments for 379 individuals.

**Table A-2 Results from model predicting PHQ-9 score before session from baseline features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.69 | [ 0.63 ; 0.75 ] | <0.001 |
| ***Number of sessions*** | -0.45 | [ -0.51 ; -0.39 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 0.96 | [ -0.01 ; 1.90 ] | <0.001 |
| ***Step group 3+*** | 3.59 | [ 2.39 ; 4.79 ] | |
| ***Constant*** | 2.30 | [ 1.31 ; 3.27 ] | <0.001 |

The results of this model (see Table A 2) suggest that baseline PHQ-9 score, the number of appointments to date and the level of severity (indicated by the Step Group) of an individual's mental health condition were all significant predictors of PHQ-9 score associated with a given appointment. The coefficient associated with baseline PHQ-9 was 0.7, meaning that every unit on an individual's baseline PHQ-9 score was associated with 0.7 of a point on the PHQ-9 outcome score on average. The number of appointments was negatively associated with PHQ-9 score, suggesting that for every appointment made, the outcome score was, on average, 0.45 points lower. Finally, the Step variables included were dummy variables comparing the association between Step groups 3 and 3+ and PHQ-9 score and Step group 2 and PHQ-9 score. This means that, all other variables being equal, an individual allocated to Step 3, was likely to have a PHQ-9 outcome score an average of 0.95 of a point higher than an individual allocated to Step 2 and an individual allocated to Step 3+ is likely to have a PHQ-9 score an average of 3.58 points higher than an individual allocated to Step 2. Diagnostic group, age group and gender were not significant in this model.

When the equivalent model was developed using only data from individuals who had completed their course of treatment and agreed discharge with their therapist, the same predictor variables were included with some changes in coefficient strength as detailed in Table A-3. This model was fitted on data from 1353 appointments involving 206 patients. The association with baseline PHQ-9 score was slightly weaker and the association with the number of appointments to date was slightly stronger. Similarly, the coefficients associated with the Step 3 and Step 3+ dummy variables were approximately 0.5 higher in both cases meaning that the difference between the outcome scores of those allocated to Step 3 and Step 2 and Step 3+ and Step 2 was, on average, 0.5 greater when considering only individuals who completed treatment as opposed to all patient cases. Taking the example of Step 3+, if all other variables were equal, an individual allocated to Step 3+ was likely to have a PHQ-9 score an average of 4.1 points higher at a given session than an individual allocated to Step 2.

**Table A-3 Results from model predicting PHQ-9 score before session from baseline features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.62 | [ 0.54 ; 0.71 ] | <0.001 |
| ***Number of sessions*** | -0.51 | [ -0.58 ; -0.44 ] | <0.001 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.49 | [ 0.22 ; 2.74 ] | <0.001 |
| ***Step group 3+*** | 3.12 | [ 2.52 ; 5.71 ] | |
| ***Constant*** | 2.54 | [ 1.26 ; 3.82 ] | <0.001 |

### A.2.2 Outcome 2 – GAD-7 score before session

This model looks at the GAD-7 score associated with the current appointment. Similarly to the previous model, data from the first appointment was not included in this model as the GAD-7 associated with the first appointment was used as the baseline GAD-7. The model was fitted on data from 1883 appointments involving 375 patients. The results for this model are presented in A-4

**Table A-4 Results from model predicting current GAD-7 score from baseline features**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.62 | [ 0.54 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.72 | [ -0.91 ;  -0.52 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.01 | [ 0.001 ; 0.03 ] | 0.040 |
| ***Diagnostic group1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.65 | [ -0.24 ; 1.54 ] | 0.091 |
| ***Diagnostic group 3 (Mixed)*** | 1.09 | [0.09 ; 2.07 ] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.13 | [ 0.15 ; 2.11 ] | <0.001 |
| ***Step group 3+*** | 3.65 | [ 2.42 ; 4.89 ] | |
| ***Constant*** | 2.86 | [ -1.59 ; 2.04 ] | 0.808 |

The significant variables and direction of association with the GAD-7 outcome score were similar to those presented in the previous model. Two additional predictors were significantly associated with outcome in this model and these were the squared number of appointments, included to consider a non-linear relationship between outcome score and time, and the diagnostic group the patient was allocated to. In the case of the diagnostic group, the depression diagnosis group was the reference group, with anxiety diagnoses and mixed or other diagnoses being compared to this. The results suggest that, with all other predictors being equal, an individual in the anxiety diagnostic group was likely to have a GAD-7 score 0.65 (95% CI = [ -0.24 ; 1.54 ]) points higher, on average, than an individual in the depression diagnostic group and that an individual in the mixed or other diagnosis group was likely to have a GAD-7 score 1.09 (95% CI = [0.09 ; 2.07 ]) points higher, on average, than an individual in the depression-based group. The squared number of sessions variable can be described as considering whether there is a non-linear effect of time (number of sessions) on the outcome score. In this case, the positive value of this predictor suggests that the negative association between the number of sessions and the GAD-7 score gets very slightly weaker as the number of sessions increased. The other predictors included in this model behaved similarly to the way they did in the previous model. The baseline GAD-7 score was positively associated with the GAD-7 score with a coefficient of 0.62 (95% CI = [ 0.54 ; 0.69 ], *p* <0.001) and the difference in GAD-7 outcome score associated with the patient allocation to Step 3 or 3+ as compared to Step 2 was also very similar.

When the equivalent model was fitted on data from only those individuals who completed their course of treatment, the same variables were significantly associated with outcome, with very similar coefficients with the exception of gender. Gender was not significant when the model was fitted on all data but was significant in the complete cases version of the model, with a coefficient of 1.29 (95% CI = [ 0.13 ; 2.45 ], *p* = 0.029). In this case, the reference group is male. The 1.28 coefficient associated with the dummy variable suggests that female patients were likely to have a GAD-7 score that was, on average, 1.28 points higher than male patients. This model was fitted on data from 1322 appointments involving 201 patients. Results for this model can be found in Table A 7.

**Table A-5 Results from model predicting GAD-7 score before session from baseline features – completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.57 | [ 0.46 ; 0.67 ] | <0.001 |
| ***Number of sessions*** | -0.76 | [ -0.99 ; -0.53] | <0.001 |
| ***Number of sessions (squared)*** | 0.01 | [ -0.003 ; 0.03] | 0.110 |
| ***Diagnostic group 1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.36 | [ 0.18 ; 2.55 | 0.041 |
| ***Diagnostic group 3 (Mixed)*** | 1.47 | [ 0.13 ; 2.80 ] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.07 | [-0.23 ; 2.37 ] | <0.001 |
| ***Step group 3+*** | 3.27 | [ 1.61 ; 4.94 ] | |
| ***Gender*** | 1.29 | [ 0.13 ; 2.45 ] | 0.029 |
| ***Constant*** | 2.86 | [ -1.59 ; 2.04 ] | 0.808 |

### A.2.3 Outcome 3 – PHQ-9 score before next session

This model looks at the PHQ-9 score associated with the following appointment. All cases within the development set were included in this as the baseline PHQ-9 and outcome scores did not overlap. This model was fitted on data from 1832 appointments from 376 individual patients. The results of the model can be found in Table A-6.

**Table A-6 Results from fixed effects model predicting PHQ-9 score before next session from baseline features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.68 | [ 0.62 ; 0.74 ] | <0.001 |
| ***Number of sessions*** | -0.58 | [ -0.77 ; -0.39 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ -0.001 ; 0.03 ] | 0.061 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 0.93 | [ -0.01 ; 1.88 ] | <0.001 |
| ***Step group 3+*** | 3.54 | [ 2.34 ; 4.73 ] | |
| ***Constant*** | 2.30 | [ 1.28 ; 3.31 ] | <0.001 |

The structure of the model was very similar to that predicting PHQ-9 score measured just before a given appointment (Table A 2) with the addition of the squared number of sessions as a significant predictor. The associated coefficient of 0.02 suggests that the effect of the number of sessions predictor on the outcome variable became slightly weaker over time. The coefficient for the number of sessions was -0.58 (95% CI = [ -0.77 ; -0.39 ], *p* <0.001) suggesting that for every appointment made to date, the PHQ-9 score associated with the next appointment was, on average, 0.58 points lower. The association between baseline PHQ-9 score and the PHQ-9 score before the next appointment was much the same as in model in Table A 2 with every point on the baseline score being associated with, on average, 0.58 of a point on the PHQ-9 score at the following session. The coefficients associated with Step group allocation are almost identical to those in Table A2 that were previously described.

When the equivalent model was fitted using only data from those who completed their course of therapy, the structure of the model was very similar with some small changes to coefficient values, but not to the direction of association. The data was fitted on data from 1286 appointments involving 204 patients. The full model can be found in Table A 7

**Table A 7 Results from model predicting PHQ-9 score before session from baseline features – completed cases only**

| ***Fixed-effects*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.61 | [ 0.53 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.64 | [ -0.87; -0.41 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ -0.003 ; 0.03 ] | 0.111 |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.46 | [ 0.20 ; 2.73 ] | <0.001 |
| ***Step group 3+*** | 4.02 | [ 2.43 ; 5.60 ] | |
| ***Constant*** | 2.55 | [ 1.23 ; 3.88 ] | <0.001 |

### A.2.4 Outcome 4 – GAD-7 score before next session

This model looked at the GAD-7 score associated with the following appointment, reported up to two days before the next appointment. This model used data from 1825 appointments from 372 individual patients.

**Table A-8 Results from fixed effects model predicting GAD-7 score before next session from baseline features**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.60 | [ 0.52 ; 0.66 ] | <0.001 |
| ***Number of sessions*** | -0.75 | [ -0.92 ; -0.58 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.007 ; 0.04 ] | 0.004 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 0.59 | [ -0.30 ; 1.47 ] | 0.113 |
| ***Diagnostic group 3 (Mixed)*** | 1.04 | [0.09 ; 2.07 ] | |
| ***Step group 2 – Ref. group*** | | | |
| ***Step group 3*** | 1.24 | [ 0.26 ; 2.21 ] | <0.001 |
| ***Step group 3+*** | 3.70 | [ 2.49 ; 4.94 ] | |
| ***Constant*** | 2.56 | [ 1.35 ; 3.76 ] | <0.001 |

As with the two versions of PHQ-9 score, this model has an almost identical structure and composition to that predicting the GAD-7 score before the session. Step group, diagnostic group, baseline GAD-7 scores, the number of sessions and the squared number of sessions were all associated with the outcome score in the same direction and with similar coefficient sizes as was the case with GAD-7 score before the session. There are only very minor variations in coefficient values. Details of the results can be found in Table A 8.

When the equivalent model was fitted including only data from patients who had completed their course of therapy, much the same situation occurred as with the GAD-7 score before the session as an outcome score. The model was fitted on data from 1272 sessions involving 199 patients. The results can be found in Table A 9. The predictor variables as above were significantly associated with outcome, with the addition of gender. In this model, the results suggest that a female patient is likely to report a GAD-7 score before the next session that is on average 1.3 points higher than a male patient, all other predictors being equal.

**Table A 9 Results from model predicting GAD-7 score before next session from baseline features – completed cases only**

| *Predictors* | *b* | | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.54 | [ 0.44 ; 0.64 ] | <0.001 |
| ***Number of sessions*** | -0.81 | [ -1.01 ; -0.60] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.004 ; 0.04] | 0.015 |
| ***Diagnostic group 1 (Depression)*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.27 | [ 0.09 ; 2.44] | 0.060 |
| ***Diagnostic group 3 (Mixed)*** | 1.37 | [ 0.05 ; 2.70 ] | |
| ***Step group 2*** | | | |
| ***Step group 3*** | 1.26 | [-0.04 ; 2.54 ] | <0.001 |
| ***Step group 3+*** | 3.36 | [ 1.72 ; 5.00 ] | |
| ***Gender*** | 1.30 | [ 0.16 ; 2.45 ] | 0.026 |
| ***Constant*** | 1.76 | [ 0.02 ; 3.53 ] | 0.048 |

### A.2.5 Cross-validation

The models presented above were internally validation using five-fold cross validation as described in the previous chapter. Table A 10 provides a summary table of the R-squared values and calibration slopes associated with this cross-validation.

**Table A 10 Summary statistics from five-fold cross-validation of developed models**

| *Outcome* | *All or completed cases* | *Mean cross-validated $R^2$* | *Range of $R^2$* | *Calibration slope* | *Intercept* |
|---|---|---|---|---|---|
| **Outcome 1 – PHQ-9 before session** | All cases | 0.50 | [ 0.37 – 0.64] | 0.97 | 0.38 |
| | Completed only | 0.41 | [ 0.30 – 0.61] | 0.95 | 0.55 |
| **Outcome 2 – GAD-7 before session** | All cases | 0.36 | [0.28 – 0.52] | 0.93 | 0.68 |
| | Completed only | 0.32 | [ 0.26 – 0.39] | 0.93 | 0.80 |
| **Outcome 3 – PHQ-9 before next session** | All cases | 0.47 | [0.34 – 0.62] | 0.96 | 0.50 |
| | Completed only | 0.38 | [ 0.24 – 0. 61] | 0.93 | 0.82 |
| **Outcome 4 – GAD-7 score before next session** | All cases | 0.34 | [ 0.23 – 0.49] | 0.92 | 0.91 |
| | Completed only | 0.30 | [ 0.20 – 0.37] | 0.91 | 1.03 |

All calibration slope scores were above 0.90, suggesting acceptable calibration of the model. The strongest model in terms of the reported mean R-squared, was that predicting the PHQ-9 score reported before a therapy session. A mean R-squared of 0.50, suggesting that it explains 50% of the variation in PHQ-9 scores, puts forward a strong model. The model predicting PHQ-9 score reported prior to the following session reported a slightly weaker mean R-squared. In the case of models predicting GAD-7 outcome scores, the associated R-squared measures were on average 0.10 weaker that those associated with the equivalent PHQ-9 models. Similarly, for each outcome put forward, the model developed using the full data set seemed slightly stronger than that developed using only data from patients who completed their course of treatment. This may, however, be associated with the number of data points included.

## A.3 Linear regression models results

### A.3.1 Outcome 5 – Final PHQ-9 score

This model considers the demographic and baseline features available in the data and how these were associated with the final outcome score reported at the end of treatment. Baseline PHQ-9, age group, step group, number of sessions attended, broad diagnostic group and gender were considered as candidate predictors in this model with final PHQ-9 score as the outcome. The model was fitted on data from 207 patient cases.

**Table A 11 Results of regression predicting final PHQ-9 score from baseline features.**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ-9*** | 0.49 | [ 0.39 ; 0.60 ] | <0.001 |
| ***Constant*** | 1.09 | [ -0.34 ; 2.51 ] | 0.133 |

The results of the model (Table A 11) suggest that only baseline PHQ-9 score was significantly associated with the final PHQ-9 score with a coefficient of 0.49 (95% CI = [ 0.39 ; 0.60 ], $p$< 0.001). This suggests that each point on the baseline PHQ-9 score was, on average, associated with 0.49 of a point on the final PHQ-9 score. The other demographic variables tested were not significant. This baseline model was estimated to explain 28.8% of the variation in the outcome data. Given that only one variable was significant in the model, this puts baseline PHQ-9 score forward as a very strong predictor of final PHQ-9 score.

### A.3.2 Outcome 6 – Final GAD-7 score

This model considered the demographic and baseline features available in the data and how these were associated with the final outcome score reported at the end of treatment. Baseline GAD-7 score, age group, step group, number of sessions attended, broad diagnostic group and gender were considered as candidate predictors in this model with final GAD-7 score as the outcome. This model was fitted on data from 203 patient cases.

**Table A-12 Results of regression predicting final GAD-7 score from baseline features.**

| ***Final GAD-7 score*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.43 | [ 0.31 ; 0.54 ] | <0.001 |
| ***Constant*** | 1.04 | [ -0.45 ; 2.57 ] | 0.184 |

The results of the model (Table A 12) suggest that only baseline GAD-7 score was significantly associated with the final GAD-7 score with a coefficient of 0.43 (95% CI = [0.31 ; 0.54 ], $p< 0.001$). This suggests that each point on the baseline GAD-7 score was, on average, associated with 0.43 of a point on the final GAD-7 score. The other demographic variables tested were not significant. This model was estimated to explain 20.4% of the variation in the outcome data when including the baseline score variable alone. As was the case for final PHQ-9 score, this is a reasonably strong baseline model and puts the baseline score forward as a strong standalone predictor of final GAD-7 score.

### A.4 Overview of results

Throughout the models presented and across the different versions of outcome scores, three predictor variables were consistently significant and with similar coefficients estimated. These were the baseline outcome score (PHQ-9 or GAD-7, model dependant), the number of appointments to date and the step group to which an individual was allocated, indicating the severity of their condition. The baseline score was positively associated with the outcome score in all models, suggesting that a higher baseline score was associated with a higher outcome score. The number of sessions was negatively associated with outcome score suggesting that the more appointments or sessions a patient made, the lower their PHQ-9 or GAD-7 score was. The Step group variable was categorical and its effect in the model was determined using dummy variables. Across the models presented, Step Group 3 was associated with outcome scores approximately one point higher than Step group 2 and Step group 3+ was associated with outcome scores 3 to 4 points higher than Step Group 2. These features were also consistent across both versions of each outcome score. This is most likely associated with the fact that only the number of session variables actually changes at different time points within the same individual, all other predictors included remained stable.

In addition to these variables that were consistently significant across the models tested, three further predictors were significant in some but not all of the models developed. A squared time variable, the squared number of appointments made to date, was included in order to test a non-linear relationship of outcome score with this measure of time. This predictor was significant in all models with the exception of that considering PHQ-9 score before a session as the outcome score. The categorical variable indicating the broad diagnostic group (Depression, Anxiety, mixed and other) was significant in all models looking at a version of the GAD-7 score as the outcome variable. The results suggested that, all other predictors being equal, an individual with an anxiety-based diagnosis was expected to have a GAD-7 score (before the current or next session) that is between 0.5 and 1.3 points higher, on average, than an individual with a depression-based diagnosis, which is consistent with what would be expected. The final variable to cover here is gender. Gender was only significant in two of the eight models considered (not all are presented), this was in models looking at both versions of the GAD-7 score when these were fitted only on data from individuals who completed their course of therapy. In this case the models suggest that being female was associated with an outcome score that was 1.3 points higher, on average, as compared to an equivalent male patient.

# Appendix B - Additional results tables

## B.1 Chapter 4 - Outcome 4 – Model of GAD-7 score at following session from LIWC categories fitted on data from completed cases

**Table B-1 Results from model predicting GAD-7 score at following session from LIWC categories – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline GAD-7*** | 0.53 | [ 0.43 ; 0.63 ] | <0.001 |
| ***Number of sessions*** | -0.82 | [ -1.03 ; -0.61 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.03 | [ 0.008 ; 0.04 ] | 0.005 |
| ***Diagnostic group 1 (Depression) – ref. group*** | | | |
| ***Diagnostic group 2 (Anxiety)*** | 1.13 | [ -0.008 ; 2.37 ] | 0.072 |
| ***Diagnostic group 3 (Mixed)*** | 1.34 | [ 0.06 ; 2.63 ] | |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.30 | [ 0.03 ; 2.56 ] | <0.001 |
| ***Step group 3+*** | 3.37 | [ 1.78 ; 4.95 ] | |
| ***Patient negations (LIWC)*** | 0.22 | [ -0.009 ; 0.45 ] | 0.060 |
| ***Therapist Negative language (LIWC)*** | 0.31 | [0.09 ; 0.52 ] | 0.004 |
| ***Therapist Positive language (LIWC)*** | -0.14 | [ -0.29 ; 0.006 ] | 0.061 |
| ***Constant*** | 2.40 | [ 0.52 ; 4.28 ] | 0.012 |

## B.2 Chapter 5 - Outcome 1 – Model predicting PHQ-9 at session from LIWC-based I2E variables fitted on data from completed cases

**Table B-2 Results from model predicting PHQ-9 score at session from LIWC-based I2E variables – completed cases only**

| *Predictors* | *b* | *95% CI* | *P* |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.62 | [ 0.54 ; 0.70 ] | <0.001 |
| ***Number of sessions*** | -0.44 | [ -0.51 ; -0.36] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.38 | [ 0.15 ; 2.62 ] | <0.001 |
| ***Step group 3+*** | 3.79 | [ 2.23 ; 5.34 ] | |
| ***Patient negative language (LIWC-based I2E query)*** | 0.29 | [ -0.005; 0.59 ] | 0.054 |
| ***Patient positive language (LIWC-based I2E query)*** | -0.40 | [ -0.66 ; -0.24 ] | 0.002 |
| ***Constant*** | 2.79 | [ 1.31 ; 4.26 ] | <0.001 |

## B.3 Chapter 5 – Outcome 3 – Model of PHQ-9 score at next session from LIWC-based I2E variables fitted on data from completed cases

**Table B-3 Results from model predicting PHQ-9 score at the next session from LIWC-based I2E variables – compelted cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.61 | [ 0.52 ; 0.69 ] | <0.001 |
| ***Number of sessions*** | -0.69 | [ -0.92 ; -0.46 ] | <0.001 |
| ***Number of sessions (squared)*** | 0.02 | [ 0.002 ; 0.04 ] | 0.030 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.42 | [ 0.17 ; 2.67 ] | <0.001 |
| ***Step group 3+*** | 3.80 | [ 2.24 ; 6.35 ] | |
| ***Patient positive language (LIWC-based I2E query)*** | -0.25 | [ -0.52 ; 0.02 ] | 0.071 |
| ***Therapist positive language (LIWC-based I2E query)*** | -0.22 | [ -0.50 ; 0.05 ] | 0.114 |
| ***Constant*** | 3.69 | [ 2.17 ; 5.20 ] | <0.001 |

## B.4 Chapter 7 – Outcome 1 – Model of PHQ-9 score at session from CTS-R based variables fitted on data from completed cases

**Table B-4. Results from model prediction PHQ-9 at session from CTS-R measures - completed cases only**

| ***Predictors*** | ***b*** | ***95% CI*** | ***P*** |
|---|---|---|---|
| ***Baseline PHQ9*** | 0.63 | [ 0.55 ; 0.71 ] | <0.001 |
| ***Number of sessions*** | -0.48 | [ -0.55 ; -0.40 ] | <0.001 |
| ***Step group 2 – ref. group*** | | | |
| ***Step group 3*** | 1.43 | [ 0.17 ; 2.68 ] | <0.001 |
| ***Step group 3+*** | 3.89 | [ 2.31 ; 5.47 ] | |
| ***Agenda setting*** | -1.72 | [ -3.36 ; -0.08 ] | 0.040 |
| ***Constant*** | 2.49 | [ 1.21 ; 3.78 ] | <0.001 |