1785

The University of Georgia

ARTIFICIAL INTELLIGENCE CENTER

CASPR Research Report 2005-02

COMPUTERIZED ANALYSIS OF SALIENT ITEMS AND DISCOURSE ORGANIZATION IN SCHIZOPHRENIC PICTURE DESCRIPTIONS

Salena A. Sampson

caspr

COMPUTER ANALYSIS OF SPEECH
FOR PSYCHOLOGICAL RESEARCH

Artificial Intelligence Center
The University of Georgia
Athens, Georgia 30602-7415 U.S.A.
www.ai.uga.edu/caspr

Computerized Analysis of Salient Items and Discourse Organization In Schizophrenic Picture Descriptions

by

Salena A. Sampson

(Under the direction of Dr. William Kretzschmar)

Abstract

Computer analysis of spoken picture descriptions shows that schizophrenic patients produce descriptions which are less complete and more disorganized than those produced by healthy controls. Completeness was measured by the relative presence of salient items mentioned in the picture. Organization was measured by the number of transitions between regions in the picture. This study also shows that exposure to cannabis can produce more transitions between picture regions. Additionally, this study brings to light a methodological concern, the possible significance of picture selection in picture description tasks. Certain pictures appear to bring out different linguistic features.

Index words: Schizophrenia, Schizophrenic discourse, Cannabis, Discourse organization, Saliency, Picture descriptions, Computational psycholinguistics

Computerized Analysis of Salient Items and Discourse Organization In Schizophrenic Picture Descriptions

by

Salena A. Sampson

B.A., The University of Georgia, 2003

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Arts

Athens, Georgia

2005

Computerized Analysis of Salient Items and Discourse Organization In Schizophrenic Picture Descriptions

by

Salena A. Sampson

Approved:

Major Professor: Dr. William Kretzschmar

Committee: Dr. Michael A. Covington
Dr. Stephen Ramsay

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

Dedication

I dedicate this thesis to my parents and sisters, who have provided unwavering advice and support throughout the years.

Acknowledgments

This research was funded by a contract with GlaxoSmithKline Research and Development Ltd., Michael A. Covington, principal investigator. I would also like to thank Cecile Henquet and the rest of the team at Maastricht for gathering the data used in this study.

Additionally, I would like to thank Dr. Covington and Dr. Kretzschmar for their constant support and guidance throughout the course of developing my ideas and writing this thesis, and Dr. Ramsay for helping me to develop my programming skills and answering all sorts of related questions. Finally, I would like to thank the entire CASPR team for their support and feedback, especially Cati Brown and Congzhou He for data management and statistical advice, Eric Morris for helping with technical questions, and Matt Voss for his continual help and support – technical, intellectual, and personal. Thank you all. Without your help, this thesis would not have been possible.

Table of Contents

Page

Chapter

Appendix

List of Figures

Chapter 1

Introduction

1.1 Overview

Computer analysis of spoken picture descriptions shows that schizophrenic patients produce descriptions which are less complete and more disorganized than those produced by healthy controls. This study follows up a previously proposed method for ranking the completeness of picture descriptions by computer, the Salient Items Test (Covington 2004). It additionally proposes a computerized method, adapted from join count statistics, which may be useful in ranking discourse organization in picture descriptions. Both of these methods and all of the data in this study are picture dependent. As such, this study gives particular consideration to the significance of picture as variable in terms of language data elicited and methodology for the computerized measurement and ranking of this data.

1.2 Introduction

Picture descriptions used for diagnostic and research purposes have a long history, though not all pictures are equally well suited for every kind of study. The Thematic Apperception Test, a popular picture description test, includes more ambigous pictures which work well for projective personality testing, but not so well for identification of concrete salient items. Most of the images were created by artists, and with no diagnostic intentions. For instance, Card 2 from the 1943/1971 edition of the TAT pictures is a black and white rendering of a painting by Leon Kroll entitled “Morning on the Cape.” Card 9BM from this same edition is an adaptation of a photograph entitled “Siesta” by Ulric Meisel. This set of pictures was selected largely because of their emotive properties. Where there have been modifications from the original source material, these modifications seem to favor more ambiguity and less detail. Card 9BM, a picture of several men lying in a field, is somewhat more ambiguous than its source photograph in that details such as chaps and cowboy hats are removed, making the identity of the men and their situation less concrete. Having seen three revised editions, the TAT pictures are still quite popular for picture description tasks in psychological research (Morgan 1995). They are rivaled in popularity perhaps only by the famous "cookie theft" picture from the Boston Diagnostic Aphasia Examination, first available in 1972. The goal of this test is to diagnose aphasic patients with a clear variety of aphasia such as Wernicke, Broca, or Conductive aphasia. Despite the special purpose of this test, the "cookie theft" picture has also become a prominent choice for picture description tasks, especially in linguistic research (Goodglass 2005).

Rather than using more canonical pictures, this study uses pictures designed specifically for the Salient Items Test to get a closer look at a couple of the classic symptoms of schizophrenia, derailment and disorganized discourse. The pictures were commissioned by GlaxoSmithKline Research and Development Ltd. and drawn by Melody Covington in 2002. They were designed to have "easily recognizable representations of objects, plants, animals, people, and/or activities" and to have a "clear interpretation" of these items and activities (Covington 2004). By using these picture descriptions to elicit language data, we are able to measure the degree to which a given subject stays focused and on task by counting the number of references he or she makes to salient items in the picture. Then by dividing the picture into discreet regions, one is able to measure discourse organization as a function of number of transitions between the regions. In data elicited from one picture, I found that schizophrenic patients had a tendency to mention fewer of the salient items. In data elicited by a second picture, I also found that patients made more transitions between regions. This may indicate a choppier or more disorganized discourse.

Another important methodological finding associated with this type of testing is that all results in picture descriptions are picture dependent. What picture is used to elicit speech is an important factor in what type of speech is produced. Not only does the lexicon vary from picture to picture, as would be expected, but also the degree to which there is a central lexicon for a picture. Comparing descriptions from two different pictures suggests that certain pictures may have a more unified lexicon. Additionally, certain pictures may produce more salient items than others, even if, as in this case, the pictures being used were designed to have the same number of salient items. Likewise, the spatial organization of the items in the picture may make it more or less difficult to measure discourse organization in the picture descriptions it produces. From these findings, one may conclude that picture descriptions are a useful tool for studying derailment and disordered discourse in schizophrenic speech, but that certain pictures are much better suited for eliciting different types of linguistic data useful for study of derailment and disordered discourse.

Likewise, certain pictures tend to produce descriptions that are more easily analyzed by computer. This is a particularly important consideration when selecting a picture, especially for the measurement of discourse organization. Pictures must have a simple overall spatial organization with clear, discreet regions, such that transitions between regions can be easily counted. While a human rater might be able to recognize disorganization in any discourse type, regardless of the organization of a picture in picture descriptions, a computer requires that one map the organization of a picture onto the expected organization of the picture description. A more clearly organized picture produces a more clearly projected organization. If one can identify discreet regions associated with particular items, one might expect items within the same region in the picture to show up more often together in the description. Though this same method could be easily implemented by hand, computer rating adds an important dimension to the task of selecting a picture, as certain pictures might be not only ideally suited to studying salient items or discourse organization, but also better suited to computer analysis. The rewards associated with the computerized study of schizophrenic research, however, far outweigh these extra considerations. With appropriate picture selection, simple computer ranking of descriptions can be quick and reliable.

1.3 Literature Review

While literature on schizophrenia and language abounds, research on disorganization syndrome, particularly as it relates to discourse organization of picture descriptions is still relatively limited. To complicate the matter, schizophrenia is recognized as having any number of manifestations. There is no specific set of symptoms that is suffered by all patients. Instead, patients must be divided into thought disordered and non-thought disordered categories, and those suffering from positive symptoms and those suffering from negative ones. Even these distinctions cannot fully account for all of the variability in symptoms suffered by people diagnosed with schizophrenia. This makes studying schizophrenia difficult in general as one cannot be certain of uniformity in patient populations or similarities between the patients in one study and those in another. In trying to study schizophrenia, one must first be careful to distinguish between patients who suffer from positive symptoms, hearing voices and having delusions, and those who suffer from negative symptoms, speaking little and having a blank affect. McKenna and Oh argue in particular that "poverty of speech was associated with negative symptoms and poverty of content with disorganization in a factor analytic study" (44), poverty of speech being a simple lack of speech production and poverty of content including individuals with fluent speech that is marked by repetition or 'empty philosophizing.'

Liddle (1987) proposes a three syndrome model of schizophrenia based on a factor analysis of classic symptoms suffered by patients. In addition to the negative-positive symptoms distinction, Liddle proposes a third syndrome, disorganization syndrome. This syndrome is defined by inappropriate affect, poverty of content of speech, tangentiality, derailment, pressure of speech, and distractability. Inappropriate affect refers to unusual demeanor and facial expressions. Pressure of speech refers to the number of words uttered per minute, including individuals who talk excessively quickly and excitedly. Tangentiality, derailment, and distractability each refer to a patient's ability to stay focused on the topic at hand. Tangentiality is replying to questions and discussion in an irrelevant manner. Distractability is measured by abrupt stops in conversation, with attention shifting to some other external stimulus. Derailment, in contrast, refers to a patient's slowly slipping off topic as a result of associating language internal structures and ideas. Each of these features contributes to what Liddle describes as disorganization syndrome, and any number of these could produce a decrease in number of salient items mentioned or a more disorganized discourse in a picture description.

Specifically in terms of picture description tasks and content, Goren (1996), using the "cookie theft" picture to elicit responses, notes poverty of content in both schizophrenic responders and non-responders to medication. He also mentions non-linguistic factors such as attention and logical sequencing. Docherty (2005) continues study on attention and logical sequencing, finding that these factors do indeed correlate with schizophrenia, but that they also correlate with disorganized speech. Gernsbacher (1999) concludes that schizophrenic patients with verbose disordered discourse shift too frequently between ideas, and that this shifting is a product of lack of attention and reduced ability to lay an organizational foundation for the discourse.

Picture descriptions, which take advantage of the explicit organization of the picture, however, also make interesting discourses in which to consider organization as it relates to schizophrenia. Smith (2003), in analyzing normal picture descriptions, concludes that reduced token counts in this form of discourse as compared to story telling is a product of organizational and vocabulary constraints associated with the picture; McKenna, Oh, and McCarthy (2002) conclude that pictures might help encourage organization in schizophrenic speech, rather than imposing, providing a framework. Comparing oral and written picture descriptions produced by Alzheimer's patients, Croisile concludes that written descriptions have grammatical and semantic errors as well as more semantic intrusions, suggesting that written picture descriptions might yield better results. Clearly, there is controversy over the impact of picture description tasks on spoken and written language, especially as these tasks relate to the complexity of the discourse.

Though there has definitely been a line of study related to schizophrenia and discourse organization and completeness, the problem remains of how to study discourse organization and completeness as can be ranked by a computer. With regards to completeness of discourse, Covington (2004) proposes a method for computer rating of "completeness or incompleteness of a description" – the Salient Items Test. With this proposal, he outlines several ways of identifying salient items in pictures and counting them. Preliminary data suggests that schizophrenic patients are more likely to leave out salient items from their descriptions, whether as a product of derailment or lack of attention or other factors. The current study is designed to investigate this hypothesis further.

Lee and Kretzschmar (1993), borrowing spatial analysis techniques from geographers, propose a way that join count statistics may be used to model dialect regions, counting margins between areas with and without a given feature. For instance, in linguistic geography, one may label any given region on a map where a particular linguistic form appears as BLACK. One may then label any region where the form does not occur as WHITE. After counting the number of boundaries between BLACK and WHITE areas, one may then compare these figures to an expected figure, derived from the probability that the particular linguistic form will appear in a given area. This same method may be adapted to fit organizational studies of picture descriptions, only regions in a picture will be coded, and transitions between these regions in the discourse can then be counted.

Chapter 2

Materials and Pictures

2.1 Materials

All of the data for this study was gathered by Cecile Henquet, Lydia Krabbendam, and Jim van Os of The Department of Psychiatry and Neuropsychology in the European Graduate School of Neuroscience, and Jan Raemakers of the Department of Neurocognition at Maastricht University in the Netherlands. The experiment was approved by the human subjects ethics committee of the University of Maastricht, and the use of the data was approved by the University of Georgia's human subjects committee. This study consists of 29 schizophrenic patients, and 31 healthy controls, all native speakers of Dutch. All subjects were between the ages of 18 and 50, all within a healthy weight range and with no respiratory or cardiovascular disease, nor brain injuries. They all have a previous history of cannabis use with no complications, have no weekly use of other illicit drugs, and no alcohol abuse (as defined by more than five drinks a day). Additionally, they all signed informed consent. Schizophrenic patients must meet all these criteria, but also have been diagnosed with non-affective psychotic disorder based on the DSM-IV criteria.

Each subject, with the exception of those who left the study, produced two picture descriptions, of pictures designed specifically for this project to be equivalents of each other, a "dog" picture and a "horse" picture as described below. One description was produced after having been exposed to a marijuana cigarette, controlled dose for body weight; the other was produced after exposure to a placebo cigarette that looked and smelled the same, but without having the active drug THC. These sessions produced a total of 55 descriptions of the dog picture, and 50 descriptions of the horse picture.

2.2 Pictures

2.2.1 Why Picture Descriptions?

Picture description tasks limit some of the variation in speech samples. Subjects are constrained by the particular task as well as the particular picture, in some ways similar to activities that require subjects retell a particular story. By limiting variation, it becomes easier to create a standard lexicon, to understand associated organizational structures, and more generally to understand and recognize typical responses to a given task. What makes picture description a particularly effective activity for the measurement of derailment and disorganized discourse is the relative ease with which one can identify salient items that a subject might be expected to mention. Even more specific to task, pictures can provide a visible organizational framework, that might allow researchers to consider how subjects organize their discourse. As subject responses are shaped by the pictures they describe, the spatial organization of the pictures may allow a researcher to anticipate possible organization of discourse.

2.2.2 The Woman with Dog Picture

In the foreground of this picture, as seen in figure 2.1, a woman walks her dog down a little path. The woman is wearing a sweater, skirt, and clogs; she is holding an umbrella. Small lines indicate raindrops, and there is little puddle behind the woman. The dog is wearing an unusual harness, and it has just stepped off the path to smell some tulips. In the background, a bus appears to be parked in front of a large cross-shaped building with a steeple that looks like a church. Not much detail is visible in either of these items, though one can see the suggestion of stained glass windows and a clock on the church.

A number of features contribute to the overall organization of this picture. First, there are a number of concrete functional relationships between items in the picture. For instance, the woman walks the dog; the dog sniffs the tulips. The woman holds an umbrella, and there are

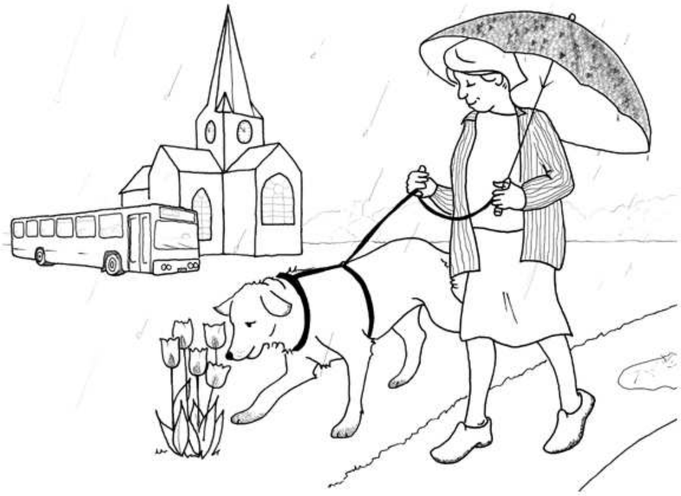

Figure 2.1: The woman with dog picture

visible raindrops and a puddle. With some overlap, there are also spatial relationships: the woman is in front of the puddle. The bus is next to the church. One particularly important observation is that items in foreground do not interact with items in the background, either in terms of action or in terms of space. A clear white space divides the two regions, and the items in the foreground bear no apparent relationship to the items in the background. This picture is defined by its discrete salient items and its discrete spatial regions.

2.2.3 The Man on Horse Picture

In the foreground of this picture, as seen in figure 2.2, a man, riding a horse, has paused at the edge of a cliff and appears to be waving at something in the distance. The man is wearing a long coat, riding pants and boots. The horse is white and has one leg lifted. In a valley below, a train is passing through, and appears to move past a cactus and some

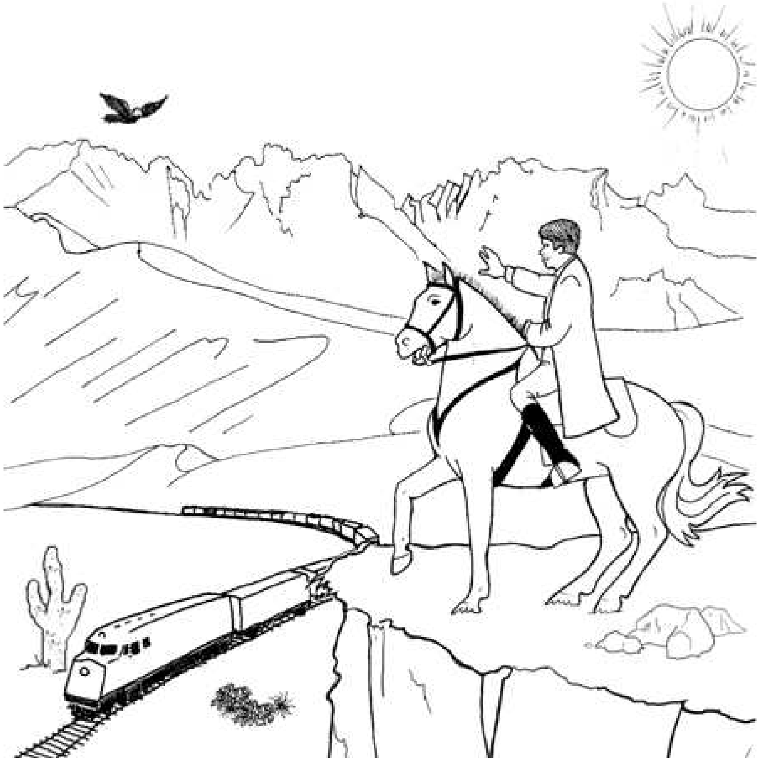

Figure 2.2: The man on horse picture

bushes. Further back in the picture is a mountainous landscape. Above the mountains, along the skyline, a bird is flying on the left side of the picture, and a sun is shining on the right.

One remarkable feature associated with this picture is the number of ways which a viewer could divide the regions. There are no clear lines between regions such as in the woman and dog picture. Though the man on his horse might constitute one region in the foreground, the background is too wide and varied to constitute any single region. A viewer might follow the line of the cliff to divide the picture in half, left and right. One could also follow a diagonal line from the train to the man to the sun, or one could divide the picture along the mountain ridge as well, making three regions – cliff, valley, and sky. The divisions of this picture are much less clear.

Likewise, the logical relationships between items are not as strong. While the relationship between man and horse is certainly logical and apparent, it is not so obvious at what exactly the man is waving. That the man, in the foreground, is waving at something in the background also makes the regions of the picture less discrete, as this gesture sets up a salient relationship between the two different regions. Additionally, the mountains, which take up a considerable amount of space, are not a particularly discrete item, either. As the man, himself, stands on a cliff, and as the mountains themselves are such a large, sprawling item, they appear to wrap around the entire space of the picture. A combination of unclear lines by which to divide the picture, complicated by a general lack of interaction between most of the salient items, with the exception of an ambiguous interaction between items from separate regions, makes the horse picture seem less clearly arranged and organized than the dog picture.

Chapter 3

Methods

3.1 Odd Types Methods

For a very specific task, such as picture description, one might expect a relatively focused lexicon. Certain words clearly related to the picture should be common to most descriptions, exemplified in salient item related words. One might expect that some schizophrenic patients, particularly those suffering from disorganization syndrome, might have an elevated level of unexpected word types, related to derailment or tangentiality, for instance. For this study, I defined these unexpected words, or odd types, as any word type that appears in any one description but did not appear in the comparative control corpus. The following sentences provide more apparent examples of odd types from the dog and horse picture descriptions:

(1.) *De grauwe gevel, en de, het groene gras het eh, paarse pad en gele klompen van de dame met haar gestreepte jasje, doen me echter niet echt charmeren van haar.*

'The drab facade, and the, the green grass the eh, purple path and the lady's yellow clogs with her striped jacket, don't really make her look attractive to me.' [4456_1, placebo patient dog description]

(2.) *[M]isschien is het wel de bus die eh mijn zoontje terugbrengt naar school als ze in Cadier en Keer in de gymnastiekzaal les hebben gehad.*

'Maybe it is the bus that eh takes my little son back to school when they have had lessons in Cadier en Keer in the gym.' [4476_1, placebo patient dog description]

> (3.) *Ja, Lucky Luke is het in ieder geval niet, want dat is eh Jolly Jumper is iets dunner volgens mij, die is daar veel te dik voor eh, hij ziet ook niet zo ruig uit.*
>
> 'Well, it's not Lucky Luke in any case, because that is eh Jolly Jumper is a little thinner I think, he is much too fat for that eh, and he doesn't look all that rugged.' [4487_2 placebo control horse description]

In the first sentence, that a path would be purple is certainly unexpected, particularly in a black and white picture. The color term 'paarse', not appearing in any other descriptions, is then an odd type. In the second example, the subject's discourse derails as the subject shifts from mentioning the bus in the picture to discussing his son's going to gym class on a bus. In this sentence, *zoontje* 'little son' and *gymnastiekzaal* 'gym' are unexpected word types not found in other descriptions of this picture. The number of odd types might also increase if a subject begins to tell a story about the picture rather than just describing. In the third example above, proper names such as 'Lucky Luke' and 'Jolly Jumper' appear as the subject speculates on the character in the picture and begins to tell a story about him. If a speaker starts to tell a story, he may associate certain items in the picture differently than other speakers, producing more odd types. Of course, what a human recognizes as 'odd' and what a computer recognizes as 'odd' are two different things. If one considers an 'odd type' to be any word type that appears in a single description, but not in the control corpus, the size and variation in the control corpus will play a considerable role in defining what exactly is 'odd'.

To count the number of these unexpected words in a given description, I wrote a simple Python program that removes punctuation and capitalization, uses a stemmer to remove morphological endings, tokenizes the data, and then compares the word list for a single description to that of the entire placebo control corpus. For placebo control subjects, I simply assembled a control corpus made up of all the remaining placebo control descriptions. I compared odd type frequencies and average position of odd types for each group, using partial correlations between total number of odd types and status and number of odd types and condition, as well as average position of odd types and status and average position and condition, controlling for total number of tokens.

3.2 Salient Items Methods

3.2.1 Identifying Items

To study the mention of salient items in the pictures, I had to first identify what is to be considered a "salient item." I defined a salient item as the following: any noun or verb that corresponds with the picture and appears within the first hundred words in a frequency ordered word list generated from the placebo control group for that picture. First, then, I had to create frequency ordered word lists for both pictures. To do this, I simply removed all punctuation and capitalization, tokenized the list, and then sorted the list in terms of frequency. With these lists, I next identified all picture-related nouns and verbs, using the first hundred words in the frequency ordered list as an arbitrary cut off point.

Having identified all of these nouns and verbs, I then had to decide which words seemed to refer to the same item within a given picture. For instance, *bloemen* 'flowers' and *tulpen* 'tulips' were collapsed into the same item; and *vrouw* 'woman' and *mevrow* 'missus' were collapsed into one item. If words clearly referred to the same item, they were grouped together. Some cases, however, were more ambiguous. For example, with the horse picture, both *lucht* 'sky' and *vliegt* 'flying' appeared in the list of top hundred words. Closer inspection of the speech samples reveals that speakers note both that there is *een vogeltje aan de lucht* 'a little bird in the sky' [4483_2], and that *de vogel eh vliegt* 'the bird is flying' [4487_2]. A speaker would not have to mention both of these items to communicate the same information. Flight implies being in the sky, and a bird's being in the sky implies flight.

That the bird is flying, as opposed to perched on a telephone wire, seems to be salient in this picture, judging from the actual speech evidence; therefore, the item should be counted. It would be inappropriate, however, to count an incomplete response for subjects who say the bird is flying, implying the sky, but not overtly mentioning it. Likewise, it would be inappropriate to count an incomplete response for subjects who say the bird is in the sky, implying flight. To complicate the matter, there is not perfect overlap in these items: subjects note other things about the sky, as well. There are no clouds in the sky; there is a sun in the sky. Since the sky is a more general, if more vague, item in the picture, I have chosen to subsume flight related words under that item, realizing that the group of words in this category is not so natural as the words in most of the other categories for either picture.

This item brings to light one of the difficulties associated with this method: there are many ways of expressing similar ideas and items. As salient items veer away from concrete nouns towards verbs and other words that express relationships between items, the numbers of ways to express the same idea increase, and they become more difficult to count. For instance, in the horse picture, 'to ride' is a salient item expressing the relationship between the man and his horse. While any number of lexical items may indicate this relationship, one can also imagine a subject saying that there is *een ruiter op zijn paard* 'a rider on his horse' [4483_2]. Though this construction clearly implies that the rider is actually riding his horse, one could certainly not count the preposition 'op' as a blanket indication of this relationship, and 'ruiter' already counts as a different salient item. Clearly, in some cases, which words constitute an item, and what exactly that item is, are somewhat ambiguous. Particularly with verbs and items communicating relationships, anticipating and counting all of the possible forms representing an item becomes a daunting, if not impossible, task.

Ultimately, in spite of attempts to create an empirical list, a researcher must admit that the final product is still, to some degree, a judgment call. Though simple word counts cannot capture all of the possible ways of expressing a given item, using frequency ordered word lists to generate a list of salient items, we are able to capture the most common variants, those which most people will use. Expecting a residue of less frequent words, the test becomes not only a measurement of salient items mentioned but also of the typicality of any given response when compared with the average.

Having identified a list of salient items using the frequency ordered word lists, I then used *The New Routledge Dutch Dictionary* and *Webster's Online Dictionary* to identify lists of synonyms for these items. From these lists, I performed substring searches, looking for any matches that might be part of compounds and not listed in a dictionary. To do this, I wrote a short Python program to iterate through the list of all the salient items types and produce another list of words that contained any of the salient items types as a substring. I was able to identify compounds such as *regenplas* 'rain puddle' and *bergvogel* 'mountain bird'. This process ensures that compound constructions were counted, and that the words with substring matches that were counted were actually compound constructions for the specified item. Though time consuming, in preliminary research on the salient items in a given picture, with a relatively small data set, such attention in assembling a word list is not only possible, but necessary.

The final list of salient items I compiled for the dog picture is as follows: dog, church, bus, woman, umbrella, flowers, clogs, leash, rain, puddle, clock, coat, grass, path, walk, tower. The final list for the horse picture is as follows: train, horse, mountains, man, sun, sky, cactus, bird, coat, valley, wave, ride, boots.

Why Frequency Ordered Word Lists?

Looking at a particular picture, a researcher could come up with a list of salient items for that picture using any number of criteria – size of the item, its position, how clear the lines are, the relative importance of the item in the scene. All of these criteria, and certainly which criteria to use, are relatively subjective. While many people might agree on many items, the final list, if produced by this method, would still be a subjective product of what an individual researcher considers to be important in the picture. Instead, we may compile a list of salient items based on actual language data. By using a frequency ordered word list to create our salient items list, we have a more direct gauge of what actual speakers consider important in the picture.

Ideally, one would want a large test population that would allow the independent creation of an objective list of salient items; one would then be able to test this list on subsequent picture descriptions. The only complication in this data set is the relatively small number of normal descriptions for each picture, just sixteen placebo control descriptions of the dog picture and eleven of the horse picture. One would need considerably larger numbers of subjects to divide the data set, testing the second half of the control subject descriptions against the salient items list generated from the first half.

Given the constraints of this data set, I have compiled salient items lists from word lists generated from the entire population of the normal descriptions for both pictures. Though dividing the data set would be ideal, some of the effects of using frequency ordered word lists from the entire normal population may be mitigated by the striking similarities between all of the word lists. For instance, comparing word lists for the placebo control population to the exposure patient population, only one item, “leash”, appears in the top hundred words from the placebo control group that does not appear in the exposure patient list. Even the frequencies at which these words appear are quite similar: “hond” makes up about 1.7% of the total tokens in the placebo control corpus, and about 1.9% of the exposure patient corpus. “Kerk” makes up about 1.1% of the placebo control corpus, and about 1.5% of the exposure patient corpus. With this level of similarity, one could compile a salient items list from a patient group which has been exposed to cannabis or from a completely normal control group and produce essentially the same list. There should be, therefore, little concern that this method will skew the results, even given the constraints of the data set.

3.2.2 Counting Items

With a completed salient items word list, I then searched each of the text samples for tokens representing these items. To do this, I wrote a Python program that first removes all capitalization and punctuation in the file, employs a stemmer to remove morphological endings, and then tokenizes the data. Next, the program iterates through the word list previously complied for each salient item, looking to see if it can find a match in the tokenized picture description. If it finds a match, it enters that token into a list of all salient items tokens. It then saves the first position of that match word in another list specific to that item. It then continues to search for other words that might represent that particular salient item, appending the first location of each of these words onto the same list for that item. The program then sorts this list to find the first position of each salient item. If the position list contains a value, the program counts a salient item. The word corresponding with the item in the first position in this sorted list is saved as the match word. Finally, the program iterates through the list of match words, checking to see what item each word represents and adding that item to a list of salient items found, then saving the first position of that word with a salient item related variable name. The final output is the total number of salient items, the total number of tokens, the total number of salient items tokens, the average first position of a salient item, the average first position divided by the total number of tokens, and the first position for each individual salient item. I compared the results for each test group, using partial correlations between each item and status, and each item and condition, controlling for total number of tokens.

Why Tokenize?

One may wonder why the extra step of tokenization is warranted for this task. Perhaps it would be just as easy to do simple string searches rather than splitting text files into list format. One might even argue that it would be easier to find your search words with string searches, but that is the exact problem. The Dutch language, with some agglutinating properties, is rich in compound words. To cope with compound words, one strategy might be simply to count any string match as a hit for the particular search item. The problem that arises with this strategy is false positives. For example, *kerkklok* ‘church clock’ is certainly not the same item as *kerk* ‘church’, nor is *ruiterlaarzen* ‘rider boots’ the same item as *ruiter* ‘rider’. Such false positives may not seem too problematic, but even alone, they would mean that a subject who focuses only on the details of a picture, leaving out the larger items, would receive the same score as someone who mentioned all the items.

Worse yet, simple string searches may return words that are altogether wrong. For instance, the subject for description 4527_2, a description of the horse picture, mentions a *regenjas* ‘rain jacket’. Though the man wears a coat, there is certainly no indication that is coat is a rain jacket. It seems more likely that the subject has just selected a highly unusual or misleading lexical item to represent the content of the picture. In more clear cut cases, a short string may be part of any number of words that bear no relationship to the search word itself. *Toren* ‘tower’, for example, is stemmed as ‘tor’, a three letter sequence that appears in any number of completely non-related words. One would not want to count every one of those words as a match for the salient item ‘church tower’.

The problem we encounter in this situation is one of balancing precision and recall. Though we do not want to miss any of the salient items, we also do not want to count any thing that is not a salient item. The fact that Dutch morphology tends towards agglutination only complicates the issue: conceivably, a speaker could generate any number of compound words for the same item. One might consider the list of compound search words for the item “boots”. It includes variants such as *ruiterschoenen* ‘rider shoes’, *ruiterlaarzen* ‘rider boots’, *rijlaarzen* ‘ride boots’, and *paardrijlaarzen* ‘horse ride boots’. By making a finite word list rather than doing simple string searches for ‘laarzen’ or ‘schoenen’, one runs the risk of missing some of these variants.

With this particular data set, however, a more troubling problem is counting false positives. Since each text sample is so short, generally somewhere between only one hundred and five hundred words, even a small number of false positives could seriously corrupt the results. In this case, it is better to err on the side of precision, knowing that some items may not be counted, but that everything that is counted should be. Ultimately, with a well researched picture, only the most unusual variants will likely be left out of the word lists.

Why Stemming?

Each salient item may be represented by a number of lexical items. Additionally, each lexical item may have a number of word forms. Stemming, in removing morphological endings, helps reduce variability in manifestations of a given lexical item. Though stemming has definite advantages, it may not be a good idea for all studies. Sinclair (1991) points out the significance of morphology in corpus data, as different word forms may have different collocational patterns. Though morphology certainly carries important information, and *tulp* 'tulip' is certainly different than *tulpen* 'tulips' in the context of these picture descriptions, if we are interested in whether or not a subject mentions a certain item and where in the picture description that item is located, unusual morphology, though still unusual, is of less concern.

Stemming is particularly useful in working with Dutch data, as it has a somewhat richer morphology than English. It is yet more imperative that we reduce the variation in word forms, as we employ exact string matches in salient item identification. Stemming helps ensure that all of the salient items in a text, whatever form they take, are counted.

3.3 Discourse Organization Methods

To examine discourse organization, I have used the actual organization of the picture as a structural framework, the idea being that items that appear closer together in the picture should also appear closer together in the description. I have, therefore, encoded the regions in the picture using a simple foreground-background distinction as corresponds clearly with the woman and dog picture and more ambiguously with the man on horse picture. I have encoded all salient items in the foreground as belonging to region one, and all salient items in the background as belonging to region two.

The resultant list for the woman and dog picture is as follows: region one includes the dog, the woman, the umbrella, the flowers, the clogs, the leash, the puddle, the coat, the path, and the activity of walking. The bus, the church, the tower, and the clock belong to region two. Rain and grass, the two remaining salient items, belong clearly to neither region and are therefore coded separately. In the man on horse picture, the man, the horse, the coat, the boots, the activity of riding, and the activity of waving are encoded as region one; and the train, the sun, the sky, the cactus, the bird, and the valley comprise region two. The mountains which seem to extend through the picture in both regions are coded separately.

As a measure of organization in discourse, I have written a Python program to count the number of transitions between regions in a given description. After removing punctuation and capitalization, and after stemming and tokenizing the descriptions, the program iterates through the description, checking each token for membership in a complete list of all salient items word types. If a token matches, it is appended to a list of match words. The program then iterates through the list of match words to identify which salient item it represents, checking for membership in item specific word lists. If the match word corresponds with an item area one, then the number "1" is appended to a coded salient items list; if the word corresponds to an item in area two, then the number "2" is added to that list. Blanks are then added at the first and last position in the list, and each item in the coded list is joined to the next item with a hyphen. The result is that each item is counted twice, once to record the transition on the right side, and then again to record the transition on the left. The blanks in the first and last positions ensure that the very first and the very last salient items are counted twice, as the other items.

The program then counts the number of pairs representing transitions between items in different regions, '1-2' and '2-1', as compared with those that represent transitions between items in the same region, '1-1' and '2-2'. I have compared the number of transitions between regions for each test group, using partial correlations between inter-region transitions and status, and inter-region transitions and condition, controlling for the total number of transitions.

An example of how this method works is as follows. The ordered list of salient items – woman, dog, umbrella, rain, church, bus, woman – would produce the encoded list "....-1, 1-1, 1-1, 1-rain, rain-2, 2-2, 2-1, 1-.....” The pair “2-1” is counted as a transition between regions, the introductory and concluding pairs as well as all pairs encoded for the same region are counted as non-transitions, and the entries including “rain” are disregarded.

3.3.1 Why Encode Picture Regions?

Conceivably, one might try to measure the movement from any given salient item token to another, taking into account which exact item transitions to which other items. This type of measure produces a great deal of variation in potential transitions. While a transition from “umbrella” to “rain” might be more common than one from “umbrella” to “dog”, measuring these subtleties is difficult without a large control corpus. Given the relatively small sample population and the short text lengths, it becomes useful to encode a picture for distinct regions, therefore reducing such variability in the output. One counts transitions from region one to region two, for instance, instead of transitions between any individual items anywhere in the picture. Encoding regions also makes text samples easier to compare to one another: one can count and compare total number of transitions within a region to those between regions more readily than one can compare any number of specific transitions that attempt to consider the particular salient items. Encoding the picture regions simplifies results, making them easier to read and interpret.

3.4 Some Limitations of These Methods

Each of these figures, odd types, salient items, and inter-region transitions, is most useful when considered as an index. These methods rely on simple counts of words and features; as such they may miss some words or count some extra ones. For instance, in defining odd types in relation to words present in the control corpus, if the control corpus is too small, the program may report completely normal words that just happen not to appear in the small control corpus. Likewise, without a considerable amount of time researching a sufficiently large corpus of descriptions for a given picture, certain rarer words that represent salient items may not be represented in the target word list.

Likewise, the program relies on salient items tokens to mark transitions between regions. Conceivably, a given speaker could be following a normal pattern of transition – such as mentioning an item related specifically to neither region, in the circumstance of this picture, "rain" or "grass" – but have selected an unusual lexical item which the program will simply not recognize. However, if the program, relying on identification of specific items, is to err, it will more likely err in the direction of not identifying enough transition points between regions by simply not identifying enough of the salient items tokens in the first place. When treated as an index, these figures may fit into a larger constellation of observations related to a given subject.

Chapter 4

Case Studies: A Closer Look at the Language Data

4.1 The Importance of Case Studies

In designing experiments, I have attempted to tailor my methods to the actual language data. As schizophrenia represents a wide range of symptoms, and hence a considerable amount of linguistic variation, case studies can be a particularly important tool in considering the disease process. These case studies aim to get a closer look at the language through the window of salient items and discourse organization. To select cases, I have used the previously described indexes for these features. I identified descriptions whose indexes suggest full coverage of salient items and strong organization of these items in discourse, as well as descriptions whose indexes suggest problems in one or both of these areas. I have also included some of the outliers identified in the boxplots in chapter five. These case studies serve two important purposes. They help illustrate how the data itself has informed this study's methods, and they serve as concrete examples of the various linguistic phenomena measured by the programs.[1]

[1] These and all descriptions in this study come from recorded speech transcribed and translated into English by researchers at Maastricht University. These transcripts were delivered in both Dutch and English aligned phrase by phrase, and with sound recordings. Full transcripts of case study descriptions, including translations and numbered time stamps with reference to position in sound recordings, may be found in Appendix A.

4.2 Salient Items Case Studies

4.2.1 Control Description 4483_2

Description 4483_2, a description of the horse picture, was produced by a control who had not been exposed to cannabis; it makes specific reference to each of the thirteen previously identified salient items. This figure is a bit higher than the average control description, which mentions eleven of the thirteen items. Immediately, the speaker gives a specific background for the picture: *Ok, ehh. Ik zie een berglandschap voor me.* 'Ok, ehh. I see a mountainous landscape before me.' The speaker then fills in the landscape with information about the other figures in the picture, moving from the foreground back: *Ehh, vooraan op de afbeelding eh ziet ge een ruiter op zijn paard, op een eh, bergwand staan. En eh die overziet een vallei. In die vallei eh, is een spoorweg en daar rijdt momenteel een trein op.* 'Ehh, in the front of the picture eh you can see a rider on his horse, on a eh, mountain side. And eh that overlooks a valley. In that valley eh, is a railway and at the moment there's a train riding on it.'

The speaker then goes on to describe the skyline, noting the lack of clouds, the shining sun, and the bird. He is very specific in his naming of items, for instance, self correcting from a description of the setting as merely *zonnig* 'sunny' to noting more concretely that *de zon schijnt* 'the sun is shining.' In this shift, the speaker chooses between a more general assertion of his impression of the weather and a more concrete description of what has lead him to this conclusion – the actual sun drawn in the picture. One can see similar patterns in descriptions from controls and patients, the subject wanting to communicate the specific details to back up his conclusions about the picture. For instance, later in this same description, the speaker explains, *Ehh de ruiter houdt de teugls strak want het paard trekt een beetje naar achter met zijn hoofd.* 'Ehh, the rider is holding the reins tight because the horse is pulling its head back a little.' The result is a very concrete, picture oriented description with specific references to salient items.

While the subject sets up the basic framework for the picture, his description remains concrete, but general. At this point in the description, the subject has named most of the salient items – the mountains, the man, the horse, the valley, the train, the sun, the sky, and the bird – but not given much detail about any of them. Having given a general overview of the picture, the subject returns to the rider to describe more of what he is doing. He speculates on his waving, his interaction with the horse. Finally, the speaker begins to fill in details, speculating that the train might be a TGV. This movement from general to specific is an interesting point to note in terms of identification of salient items. Conceivably, any population of speakers could produce a very lengthy list of words for the item 'train'; however, if most speakers start with a more general description, even a limited word list should be able to capture the most common general lexical items for each of the salient items.

As the speaker introduces more and more detail, the organization of the discourse breaks down as the speaker begins to shift more frequently between the different regions in the picture. It is at this point that the speaker introduces the remaining salient items. For instance, the speaker concludes that *in de vallei het precies een woestijn landschap* 'in the valley it looks like a desert landscape' and explains his conclusion with the details *Er staat een cactus en wat verdorde struikjes.* 'There is a cactus and some small withered bushes.' The speaker then goes on to speculate on what type of bird might be in the sky, considering its specific color patterns, and then ultimately to describe the physical appearance of the rider in more detail, discussing his clothing, boots, and hair.

Having provided first a general description of the picture and then filling in the description with details specific to each region and item, the speaker finally begins to consider a possible story to explain the interaction in the picture. This speculation seems to be prompted by the rider's waving: *Ehh, op het eerste gezicht vraagt ge meteen af waar hij naar aan het wuiven is.* 'Ehh, at first glance you immediately wonder what he's waving at.' In the remaining seconds, the speaker concludes, *Misschien dat hij een leider is van een roversbende, om eh om over te gaan de trein overvallen.* 'Maybe he's a leader of a band of robbers, to eh about to go to raid the train.' At this point, the speaker, having exhaustively described everything in the picture, moves away from a description to more of a story telling format. This part of the description is significantly less dependent on the actual salient items in the picture, instead focusing on one detail of the picture that invites speculation, the rider's waving. It is important to note, however, that this kind of speculation does not appear until the very end of the description, after the speaker has already described all of the salient items in detail.

4.2.2 Patient Description 4498_2

Description 4498_2 of the horse picture was produced by a patient who was exposed to cannabis. At 101 tokens, this description is on the shorter end of the spectrum, but what is particularly remarkable about this description is the dramatically low number of salient items that appear in it. The program has scored only four of the possible thirteen items for this description, compared to the average eleven items. Upon closer inspection, it becomes apparent that a general lack of specificity in the description has produced this low score, perhaps even more so than the abbreviated text length.

The subject begins his description with the general comment that *Het gaat zich dus over een landschap* 'It's obviously about a landscape.' Compare this introduction with the more specific introduction from control description 4483_2, which also concerns the landscape: *Ik zie een berglandschap voor me* 'I can see a mountainous landscape before me.' One might expect that the speaker from 4498_2 is merely following the familiar general to specific pattern, which would mean that he would introduce the mountains at a later point in the description. Oddly, the speaker never mentions the mountains, however. The closest he comes is to mentioning the cliff. The generality of the first sentence largely marks the entire description.

While most speakers at some point call the person on the horse a 'boy' or a 'man' or some other more specific word, this speaker says only that *En iemand is met een paard er op uit* 'And somebody's riding on a horse.' Clearly, gender is implied by masculine pronouns used later, but the generality of introducing this central figure with only the pronoun *iemand* ‘somebody’ follows with the lack of detail in the rest of the description. As the description progresses, it becomes apparent that the speaker will provide no more detailed description of rider either, missing descriptions of items such as the coat or boots.

The description is also marked by what appears as a kind of imprecision in the lexicon as pertains to salient items and relationships. While most speakers note that it is a sunny day or that the sun is shining, the speaker of 4498_2 notes merely that *Het is vandaag een zomerse dag* ‘Today is a summery day.’ Conceivably, the subject is picking up on the same cues in the picture as the other speakers, the sun and the rays radiating from it, but is simply interpreting them in a different way. Ultimately, what has caused the speaker to conclude that it is a summery day remains unapparent, however, as he does not feel compelled to explain his conclusion with picture-related details, as many other speakers do.

Even the speaker’s attempts to explain conclusions about the picture are very general. For instance, he concludes at one point, *Maar ik denk, zoals op het plaatje te zien is dat het paard van iets schrikt* ‘But I think, as can be seen from the picture, that the horse is startled by something.’ In this sentence, the speaker seems to be aware of the need for picture related evidence to support his conclusion, but never explains precisely what in the picture makes him believe that the horse is startled. More classical associations between items are still preserved, however. For example, immediately after discussing the horse, the speaker goes on to note that *Het is een droog gebied* ‘It’s a dry region’ with the explanation *Er groeien cactussen* ‘There are cactus growing.’ As in other subjects’ descriptions, this speaker is correctly able to use cactus as evidence of an arid climate.

At points, even if the speaker notices one of the salient items in the picture, he seems to actually misinterpret it or select unusual or even incorrect words to describe it. For instance, the speaker recognizes the relationship between the man and the train, noting that *Hij wijst naar de trein* ‘He’s pointing at the train.’ With this interpretation in mind, the speaker concludes that *Misschien dat hij dat hinderlijk vindt dat daar de trein rijdt* ‘Maybe he’s annoyed by the train riding there,' in spite of what appears to be a smile on the man's face. In this circumstance, the subject has keyed in on one of the salient relationships in the picture, that between man and train; however, most subjects interpret the man's raised hand as waving. Given the usual interpretation of this gesture, the program's word list for this item includes all waving related words. One might argue that the subject's mention of pointing indicates that he has clearly noted this item; however, since the subject has misidentified the item, or at least provided a very unusual interpretation of it, this mention is not counted in his salient items total. In this description, cannabis has magnified the symptoms of schizophrenia, producing a description marked by such general lack of detail and imprecision in naming salient items.

4.2.3 Patient Descriptions 4500_1 and 4500_2

Some patients suffering from negative symptoms of schizophrenia may present with drastically reduced pressure of speech, resulting in very short descriptions. Descriptions 4500_1 and 4500_2 represent such a case. Description 4500_1 was produced by a patient who was not exposed to cannabis, and 4500_2 was produced by the same patient after exposure.

What is particularly remarkable about description 4500_1 is the number of salient items the patient mentions, in spite of the otherwise drastically reduced length of description. The entire description is only twenty five tokens total, and a surprising eight of these refer to salient items, including references to seven total items – the man, the horse, the train, the cactus, the mountains, the sun, and the bird. The syntax is drastically reduced to the point that the description reads as a list: *Ik zie een...paard met een man erop. Een zadel, halster. Een trein zie ik. Een cactus. Bergen. Hoge bergen. Een zon. Een vogeltje nog.* 'I see a...horse with a man on it. A saddle, halter. A train I see. A cactus. Mountains. High mountains. A sun. A bird as well.' The only two words in the whole description that indicate how any one item relates to another are in the phrase *een...paard met een man erop*'a horse with a man on it.' By any standards, this description is rather sparse, but most of the salient items, seven of the total possible thirteen, are preserved. Compared to the average eleven items mentioned in control descriptions, this description manages to include a relatively high number of salient items, especially given its length.

This case study becomes particularly interesting when considering the description produced when the patient was exposed to cannabis. Description 4500_1, at twenty one tokens, is yet a bit shorter than 4500_2. Additionally, the syntax becomes considerably less intelligible, and the number of salient items represented drops to only four out of a possible total of sixteen. This number is compared to the average thirteen items for patients, and fourteen items for controls. This description lacks any clear organizational framework as well as clear lexical reference: *Een bus ervoor...Bus ervoor op de grond een hond erbij. En een vrouw die de hond uitlaat. En eh...paraplu.* ‘A bus in front of it on the ground a dog there. And a woman that is walking the dog. And eh...umbrella.’ What the bus is in front of, as well as where exactly ‘there’ is, remains ambiguous. Clearly, large portions of this picture are simply not represented, and the salient items do not even appear in list format as in the previous description. Cannabis dramatically magnifies the effects of schizophrenia in this patient.

4.3 Discourse Organization Case Studies

4.3.1 Control Description 4458_1

Description 4458_1 is description of the dog picture, produced by a control who has not been exposed to cannabis. This description represents a strongly organized discourse, with only 5.88% of the transitions in the description being between regions. This figure is even a bit lower than the average percent for control descriptions, around 10%. Within the first couple of sentences, one observes the introduction of two separate types of organizational structures: movement from general to specific and an overt mention of how the speaker intends to organize the order of the items mentioned.

The subject begins the description with a very general statement which may function as a frame: *Ehh, ik zie een zwart wit plaatje.* ‘Ehh, I can see a black and white picture.’

Next the subject communicates that *Op het plaatje van links naar rechts zie ik een ehh, bloem, zie ik bloemen, tulpen.* 'On the picture from left to right I can see a ehh, flower, I can see flowers, tulips.' In this sentence, the subject informs the listener that the description of items will be moving from left to right, again setting up a basic frame for the order of the discourse. Also, in this sentence, as the subject self corrects the word he chooses for the first item, 'tulips', one may again observe the previously observed movement from general to specific. The subject begins with just the singular 'bloem', then shifts to the plural 'bloemen', ultimately specifying the particular flower 'tulpen'. The same pattern appears later in the description when the subject first notes a dog, and then specifies that *het lijkt op een golden retriever.* 'It looks like a Golden Retriever.'

In terms of transitioning between items in the picture, the subject adheres to the organizational pattern he presents at the beginning of his description, left to right movement. Interestingly, however, he maintains the foreground-background distinction, natural to this picture. The resulting movement throughout the picture is as follows: tulips-dog-woman-rain-path. The subject then returns to the dog to give a point of reference in the foreground, explaining that *achter de hond eh, staat een grote kerk.* 'Behind the dog eh, is a big church.' This sentence spatially relating items in the foreground and background serves as a transition between descriptions of the two areas.

For the description of the background, the subject provides a detailed description of the church and bus and how they are positioned in relation to each other, then moves yet further back to describe the faint lines along the horizon. Having given this level of detail, the subject regroups his thoughts *Ehh...kijken* 'Ehhm...let's see' and begins with a more detailed description moving from the background towards the foreground again, church-bus-grass-tulips-woman-dog. In this series, "grass", which is in both the foreground and the background, serves as a natural transition between items in the areas. Having exhausted most of the detail in the foreground, noting even that *Waarbij ze de rechter hand eh, de riem om haar pink laat gaan*, 'And her right hand, eh she's holding the lead around her pinky,' the subject pauses and marks yet another transition with a simple *Ja* 'Well'. At this point in the discourse, the subject begins to provide more detail about the bus, producing no more than a few words before time runs out.

One may note a number of qualities and features that help produce the overall organization of discourse in this control subject's picture description. First, at the outset, the speaker gives the listener a basic order of description, left to right movement, to help frame the picture. Also, early on, the speaker establishes a pattern of general to specific information. This activity alone might help a speaker organize his thoughts, but this movement marks not only individual sentences, *zie ik bloemen, tulpen.* 'I can see flowers, tulips,' or sections of description, *een vrouw met kort haar. Ehh, ze draagt een blouse met strepen en een rok...* 'A woman with short hair. Ehh, she's wearing a blouse with stripes,' but also the flow of the entire discourse. When the speaker returns to previous sections towards the end of his description, he attempts to add detail. For instance, he breaks the church down into *drie blokken en een toren* 'Three blocks and one tower,' or specifies that the bus is on the left side of the church, not just in front as he had previously indicated. Moving in rounds, the speaker gradually fleshes out the details of the picture.

Observing how the movement from general to specific informs the overall structure of the discourse, one might argue that this strategy allows a speaker to negotiate time and relevance. For example, if a speaker is aware of having only a certain amount to time to describe a picture of a woman, her dog, a church, and a bus, the speaker would not want to provide an overly detailed description of the woman, neglecting to mention the other relevant items. Likewise, if a speaker has given a description at a certain level of granularity, only to discover that more time remains, he may conclude that more detail was required of his description. He may therefore return to discuss a previous section of the picture with more detail. This type of negotiation of detail fit into a larger organizational framework outlined by the reader, left to right movement, and the natural organization of the picture, with distinct foreground and background areas, define the overall flow of discourse in this description.

4.3.2 Patient Description 4619_1

Description 4619_1 is a description of the dog picture, produced by a patient who has not been exposed to cannabis. Of the total transitions in this description, 20.00% are between regions, compared to the average control description with 10% inter-region transitions. Compared to the previous control description, this discourse shows some relatively early signs of disorganization. The subject simply begins his description without any indication to the listener about how his description will be moving throughout the picture: *Ik zie een mevrow. Die is haar hond uit aan het laten.* ‘I see a lady. She’s walking her dog.’

The description, however, is neither devoid of deictic words nor logical relationships between items. For instance, the subject at times describes clearly how items relate to each other spatially, noting that *Voor die hond zie ik vijf tulpen* ‘In front of the dog I can see five tulips.’ Likewise, the subject draws natural relationships between items in the picture and what they may indicate, conjecturing that the dog is likely a guide dog *want hij heeft een speciaal tuig om* ‘Because he’s wearing a special halter.’ Furthermore, after noting that it is raining, he mentions that the woman has an umbrella.

What appear to break down are the transitions between these items and ideas. For instance, the subject moves immediately from observing that *De hond is een blindengeleidehond waarschijnlijk, want hij heeft een speciaal tuig om* ‘The dog is probably a guide dog for the blind, because he’s wearing a special halter’ to *Het regent op die dag* ‘It’s raining on this day’ with no expressed transition, and no apparent correlation between these two ideas. After describing the woman’s clothes very generally, the subject transitions into describing the background. He signals that the description will be moving to the background with the prepositional phrase *Op de achtergrond* ‘In the background,’ but provides no other concrete connection between the woman’s vest, the last item he was describing, and the bus which he begins to describe. Comparatively, a normal subject may be more explicit in his transition. For instance, one might recall a strategy employed in the previous description, 4458_1, in which the speaker overtly explains that one item in the background is directly behind another in the foreground, helping to smooth the transition. Still, there are obviously many different strategies for transitioning between different parts of a description, and this speaker does signal that the listener is now to focus attention of the background. The speaker next provides a couple of sentences about the church and bus, observing a clock on the church that indicates a time of ten minutes until one o'clock. It is at this point, still relatively early on in the description, that the speaker makes a first unanticipated transition between the two discrete areas in the picture, foreground and background. Immediately after observing the time on the church clock, the speaker comments that *Het soort hond wat ik zie is eh denk ik een golden retriever.* 'The kind of dog I see is eh a Golden Retriever I think.' After this isolated comment about the dog, the speaker then immediately returns to the background to note that the bus has no clear logo on its side. After this single comment, the speaker then shifts to the foreground again, deciding *Ik kan de mevrouw beter proberen te beschrijven* 'I'd better try to describe the lady.' The span of these four sentences marks four unanticipated jumps in description between the different areas of the picture. After all of this shifting between background and foreground, the subject is finally able to focus his description on the woman and her dog for a couple of sentences, long enough to communicate the relationship between these three items – that the woman is holding the dog by a leash, and that the dog is sniffing the flowers.

Having explained the relationship between these items, the subject again shifts to the background, but this time notes the transition: *Op de achtergrond zie ik nog vaag bergen* 'In the background I can vaguely see mountains as well.' The fact that he employs such deictic words demonstrates that he is obviously aware of the different regions of the picture. The awareness of these regions simply does not structure his discourse to the extent that it might in a normal description. Throughout the remainder of the description, the subject continues to transition between foreground and background. Notably, however, as the description wears on, the length of the intervals between transitions becomes longer as the subject becomes comfortable discussing details of the description. This ability to focus on a particular item in the picture appears to coincide with a long pause in the description followed by a prompt from the experiment leader: *Is er nog meer wat je zou kunnen vertellen?* ‘Is there anything else you can tell?’ One might conjecture that at this point the speaker feels comfortable that he covered all of the basic information about the picture, and can then proceed to give more detail.

Regardless of what motivates this shift in description, though it follows a similar pattern to the previous control description, general to specific, the strategy is markedly different. Whereas the subject who produced the description 4458_1 motivates his transitions and gradually moves towards providing more detail, comparatively, the speaker of 4619_1 appears to jump about erratically throughout the first half of the description. While each time the subject shifts from one section to the next, it does appear that he is trying to add detail, much as in 4458_1; the speaker of 4619_1 is largely unable to focus on any given item, and unable to transition naturally between items, most of the detail appearing only at the end of the description.

4.3.3 Patient Description 4470_1

Description 4470_1 of the dog picture has been produced by a patient who has not been exposed to cannabis. In this description, 30.03% of the total transitions recorded by the program are between regions, as compared with the average 10% for control descriptions. This description is another example of a discourse the programs rates with a high level of transitions between different areas of the picture. Within the first sentence, the subject introduces many of the salient items: *Een vrouw wandelt op een voetpaadje met haar hond.* ‘A woman is walking on a small footpath with her dog.’ The subject clearly indicates logical relationships between items: *Het regent, zodat ze een paraplu boven haar hoofd draagt.* ‘It’s raining, so she has put up an umbrella over her head.’ The informant also clearly describes interactions between items and their spatial relationships to one another: *De hond ruikt aan wat tulpen, die naast het paadje.* ‘The dog smells some tulips, that are growing next to the small path.' The first few sentences of this description give all of the appearances of a well organized description with clear understanding and communication of relationships between items.

After describing the foreground, the subject marks his transition into the background with the introductory phrase *Op de achtergrond* 'In the background.' He notes the church and the bus, and then that the church clock reports a time of ten minutes until two o'clock, at which point he transitions back into a description of the foreground with another weather comment. While mentioning the rain is a common enough transition between foreground and background in these descriptions, as the rain is item that belongs neither to the foreground or the background, this particular weather transition is a bit unusual.

After having spent a considerable time discussing the rain and the woman's umbrella, the patient comments, *Het is erg licht uh, de zon schijnt kennelijk wel.* 'It's very bright [uh], the sun is obviously shining.' This conclusion seems somewhat remarkable given the salience of the rain and rain-related items in the picture, as well as the lack of any concrete indication of the sun. Perhaps the subject is uncertain how to interpret the seemingly white sky in the black and white picture, but these conclusions appear a bit unusual, if not somewhat illogical.

This transition provides an interesting point at which to consider how the program treats various transitions. Since one of the salient items in this picture, "rain", is not located in either discrete region of the picture, foreground or background, but is rather a weather descriptor, universal to the whole picture, the program does not assign "rain" to a particular region. The result is that transitions between any item in the foreground or background and any rain related word type are not counted as direct transitions between foreground and background items. In this circumstance, though weather words, specifically those related to rain, are a common transitioning devise in describing this picture, and would not be counted as a direct transition, this subject's use of sun related transition does count in this descriptions total transition index. Though the pattern of transition is recognizable, weather word as transitional item, the lexical item, “sun”, is an unusual weather word to select in the context of this picture.

The program picks up a similar phenomenon towards the end of this description. The speaker begins to postulate on the relationship between the woman and dog in the foreground and the church and bus in the back ground. The subject concludes his description with speculation such as, *Misschien moet de vrouw wel uh naar de bus en ziet ze dat de bus stopt, zodat ze daar naar toe kan lopen en de bus uh kan nemen naar ergens anders.* ‘Maybe the woman is on her way to the bus and she sees the bus stopping, so she can walk over there and [uh] take the bus to somewhere else.’ The speaker seems to be aware that this speculation may be peripheral information as he asks, *Moet ik blijven vertellen* ‘Should I continue?’ After a pause, and without an apparent prompts, he continues to speculate, however: *Misschien moet gaat de vrouw gewoon langs de bus naar de kerk. Ik weet niet of een hond de kerk in binnen mag.* ‘Maybe the woman must...is simply going to walk past the bus to the church. I don’t know if the dog is allowed in a church.’

The program counts each of these movements between “woman” and “bus” and “dog” and “church” as a transition, which they clearly are not. Again, the program is marking something unusual, however. This description has a high transition index, not because the speaker necessarily shifts between regions of the picture more frequently than average, but because the speaker actually sees a relationship between items from the different regions of the picture. Though nothing in the picture suggests a relationship between the woman and dog and the church and bus, but this individual spends several sentences speculating on what the possible relationship might be. Speculation of this kind is infrequent in other descriptions, if a speaker suggests a relationship between anything in the foreground and anything in the background, it is only mentioned briefly, certainly not enough to affect the picture’s transition index.

In description 4470_1, the patient seems to have no problem communicating enough of the salient relationships between items; indeed, the problem appears to be quite the opposite.

This subject finds relationships where there are none: in spite of the rain in the picture, the subject perhaps relates the white sky of the black and white picture to indications of a sunny day. Because the woman and the dog are in the picture with the church and the bus, the patient assumes there must be some correlation between them, even though none is indicated by the picture. Though the peculiarities associated with this picture description may not classically be described as disorganized discourses, the program marks each of these movements between items from the different regions. The result is that this description has a high disorganization index, if for somewhat unexpected reasons.

Chapter 5

Results

5.1 Odd Types

Knowing that schizophrenic patients may suffer from symptoms such as derailment, stilted language, and clanging, one of my earliest hypotheses was that patients might use more words that are seemingly non-related to the picture they are describing. If they did not use entirely non-related words, their lexicon might be different enough from the control group that the difference might be quantifiable. Comparing the number of odd types in each group's descriptions, I found that neither status nor condition correlated significantly with the odd types or their position.

What I did find is that the number of odd types in picture descriptions correlates with what picture is being described. Using the Wilcoxon Signed Ranks Test because of the skewness of my data, I found that horse picture descriptions have significantly higher percentages of odd types than dog picture descriptions, $p < .001$ with $Z = -3.191$. As can been seen in figure 5.1, the number of odd types increases almost linearly with the total number of types in a description. However, one can also see the overall higher levels of odd types found in horse picture descriptions. These results suggest that the lexicon for describing the horse picture is considerably less focused in all populations when compared with the lexicon for describing the dog picture.

5.2 Salient Items

Noticing the differences in the focus of the lexicon for the two pictures, one might also expect different numbers of salient items or different frequencies for these items comparing

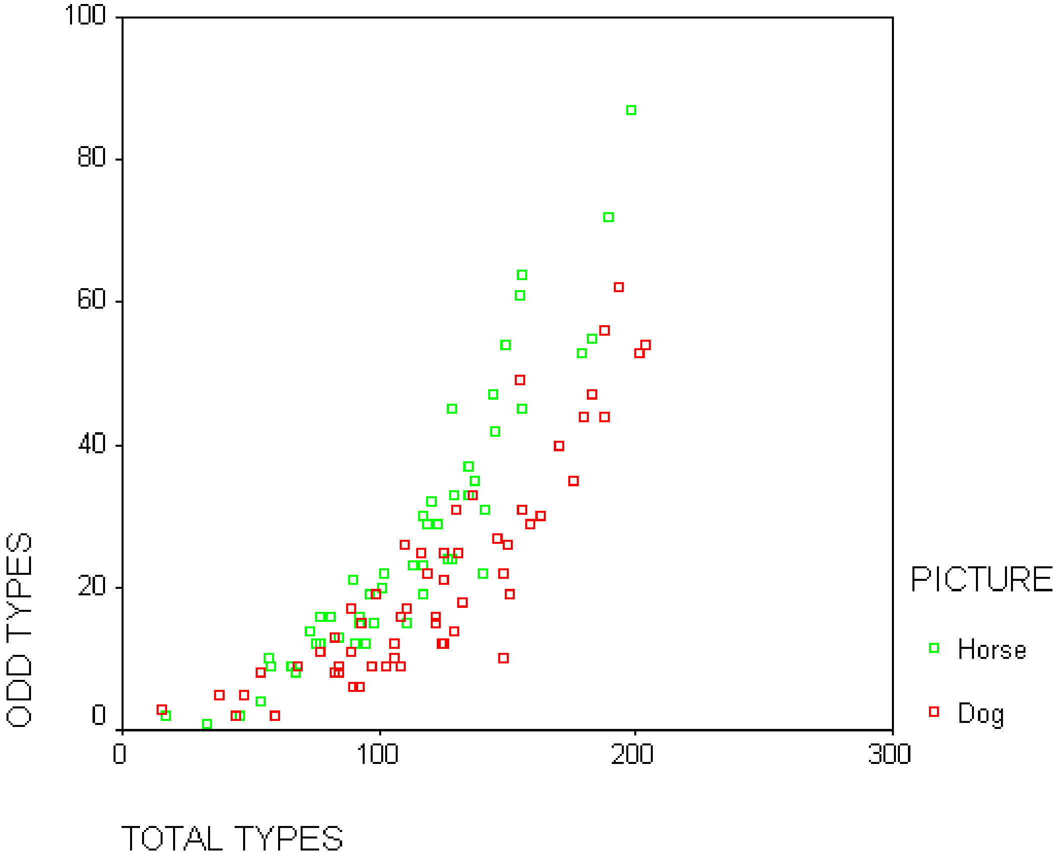

Figure 5.1: Odd types

descriptions from both pictures. Indeed, counting all of the picture related nouns and verbs for the dog picture produces sixteen salient items, whereas it produces only thirteen for the horse picture. Furthermore, looking at actual frequencies of these items reveals that while six of the sixteen salient items in the dog picture have representative words that appear at frequencies higher than 0.5% of the total tokens, only four items have corresponding word frequencies above 0.5% in the horse picture. Though these differences are subtle, they seem to correspond with the results from the odd types experiment. If salient items are measured by looking at items most subjects mention, then it stands to reason that a picture with a less centralized lexicon, such as the horse picture, might also have slightly lower frequencies of salient items or perhaps fewer salient items altogether.

Much as the pictures seem to suggest different numbers of salient items, they also produce different results in terms of significance of actual counts of salient items. With both pictures, the number of salient items tokens present in a given description correlated neither with status nor condition. Therefore, one may conclude that neither individuals who have been diagnosed with schizophrenia nor individuals who have been exposed to cannabis fail to focus on salient items to normal degree. Likewise, when considering the position of the highest frequency salient item in both pictures, 'dog' and 'train', it appears that at least for the most major items, patient and exposure groups mention them at approximately the same point in the discourse as control groups. There is no significant correlation between position of the most salient item and status or condition. This finding suggests that both patients and controls, individuals who have been exposed to cannabis and those who have not, progress to discussing central items in the picture at the same rate.

Results for both pictures start to differ more when actually looking at raw numbers of salient items mentioned. For the dog picture, the total number of salient items mentioned is not significant. For the horse picture, the number of salient items mentioned correlates with status at a level of .2430, $p < .05$. One of the major differences between the patient group and the control group in data from this picture, as seen in figure 5.5 [1], is the level of variance with the groups. Including some patients who mention all the thirteen items and others who mention as few as six within the limits of the outer quartile, the patient group demonstrates considerably more variation. This distribution might be expected due to the variety of widely different symptoms suffered by schizophrenic patients.

Though not statistically significant, one finds a similar distribution of salient items mentioned with the dog picture. In the case of the dog picture, as seen in figure 5.4, the mean number of items mentioned is slightly lower in patient descriptions, and there is slightly more

[1] For this and all subsequent graphs, circles accompanied by numbers represent descriptions that are outliers. These outliers are significant as many of them have become case studies in this thesis.

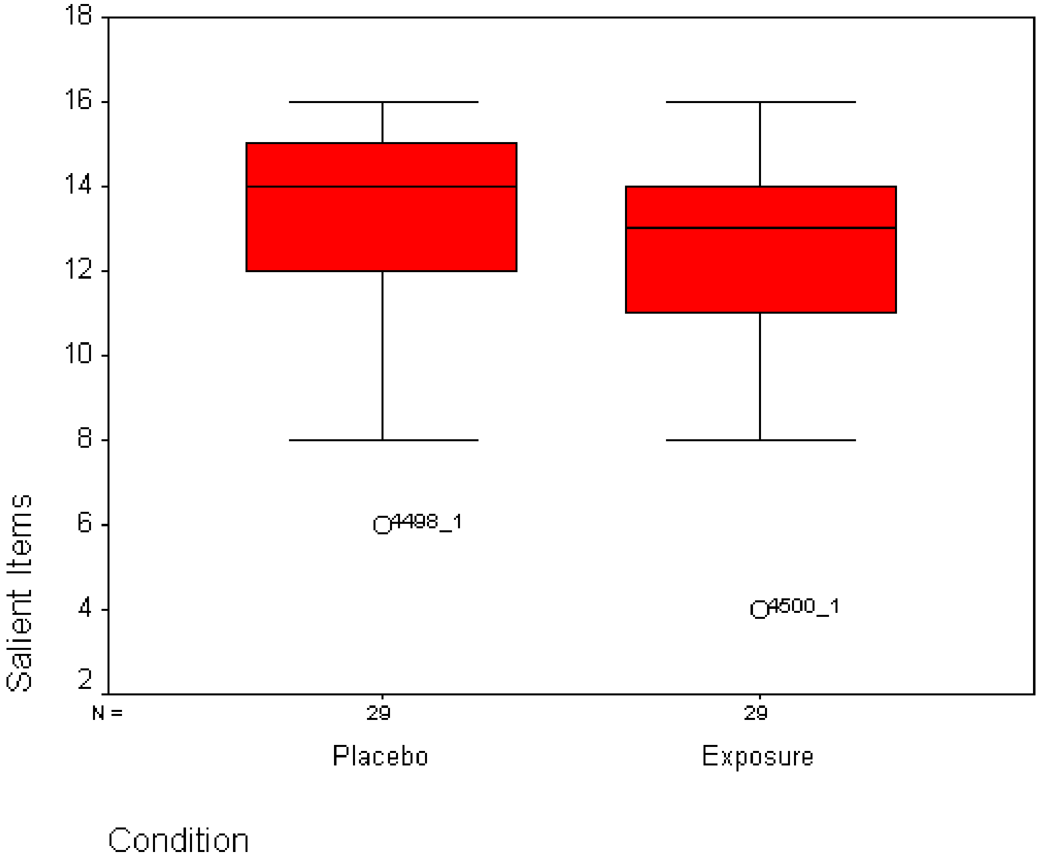

Figure 5.2: Condition as related to salient items in the dog picture

variation in the patient group. One can observe a similar phenomenon in data collected from both pictures when considering condition. In the box plot from the dog picture data, figure 5.2, while the variance is the same, the mean number of salient items mentioned is lower in the exposure group than the placebo group, thirteen items as compared to fourteen. In the results from the horse picture, as can be seen in figure 5.3, though the mean number of items mentioned is the same in both populations, there is considerably more variation in the exposure group. Though not all of these results are significant, the distribution of the data seems to suggest a consistent pattern: there is more variation in the patient group than the control group, meaning that some patients mention considerably fewer salient items than

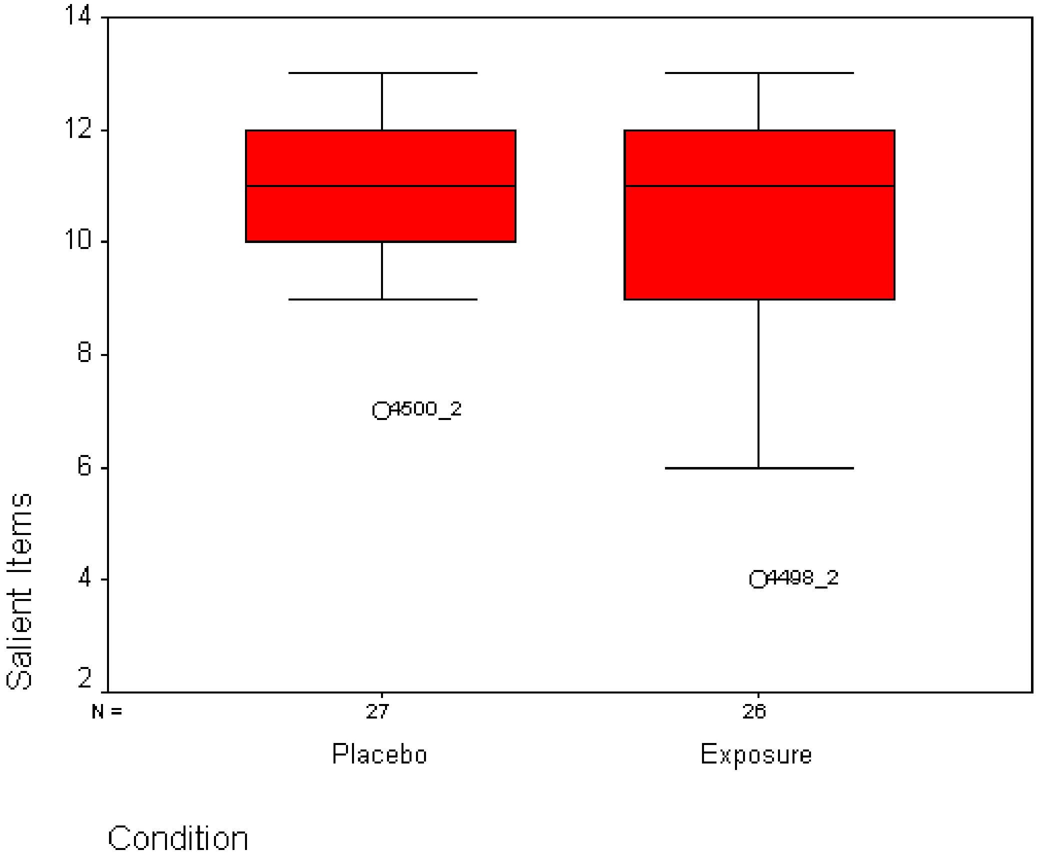

Figure 5.3: Condition as related to salient items in the horse picture

the normal population. Likewise, the population exposed to cannabis shows more variance, with more individuals with lower levels of salient items. Though a larger data set or slightly modified pictures may be needed to further test this hypothesis, counts of salient items seem to have the potential to be a useful measurement.

5.3 Discourse Organization

One witnesses similar picture-dependent results when considering discourse organization. The horse picture, the more disorganized of the two pictures with less distinct regions, produces no clear correlation between status or condition and the number of transitions

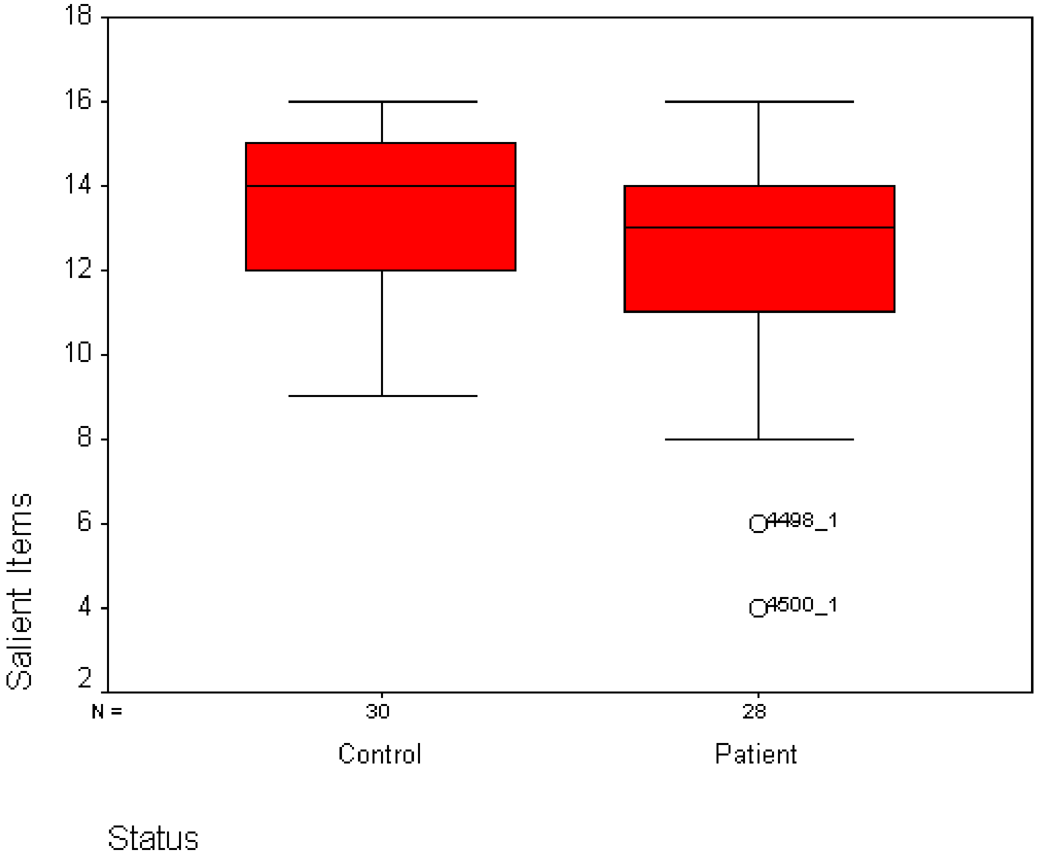

Figure 5.4: Status as related to salient items in the dog picture

between regions. The results from the dog picture, however, indicate a correlation between the number of inter-region transitions and status at a level of .2404, $p < .04$, when controlling for total number of transitions. As seen in figure 5.7, the patient group demonstrates a higher percentage of inter-region transitions and considerably more variance, including individuals with inter-regional transitions constituting over 30% of their total transitions. These results also indicate a further correlation between inter-region transitions and condition at a level of .2690, $p < .03$, also controlling for total transitions. One finds a similar pattern comparing the placebo and exposure groups to the control and patient groups, as shown in figure 5.6. The mean percent of inter-regional transitions is elevated in the exposure group, and the

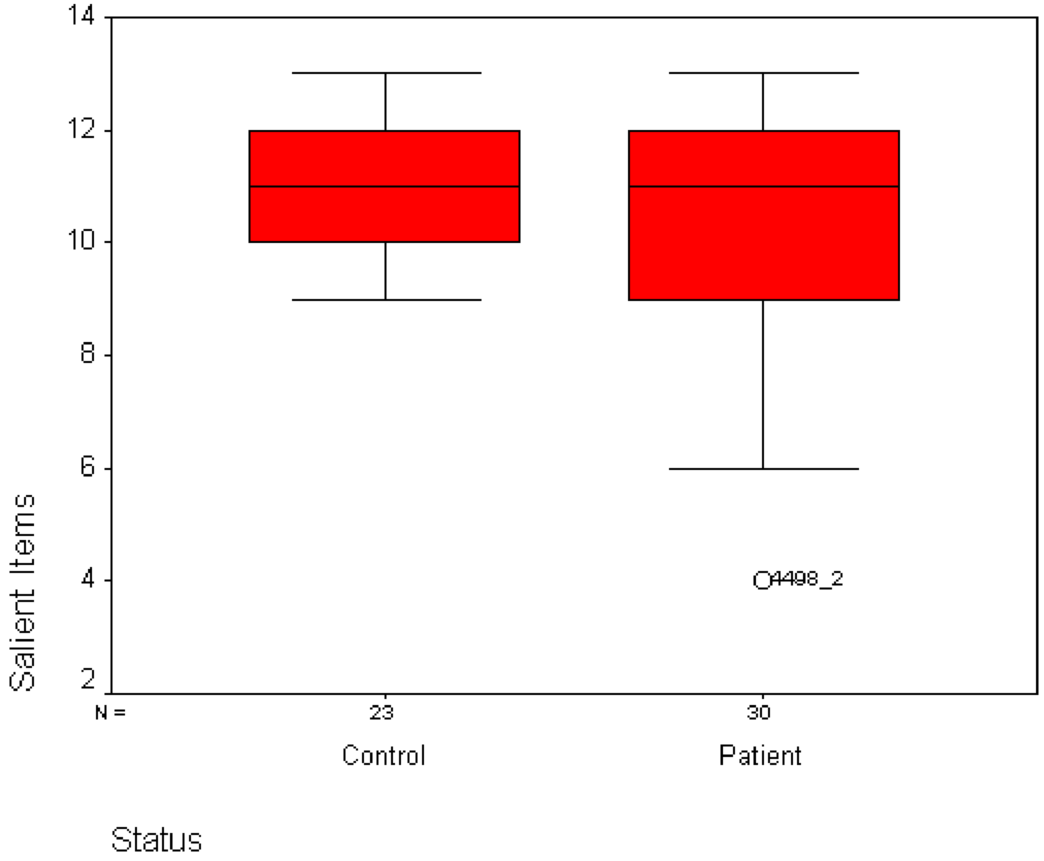

Figure 5.5: Status as related to salient items in the horse picture

exposure group also shows more variance. Both schizophrenic patients and subjects exposed to cannabis are more likely to have elevated numbers of transitions between regions in the picture, suggestive of a less organized discourse. On this level of organization, the effects of cannabis may mimic one of the symptoms of schizophrenia.

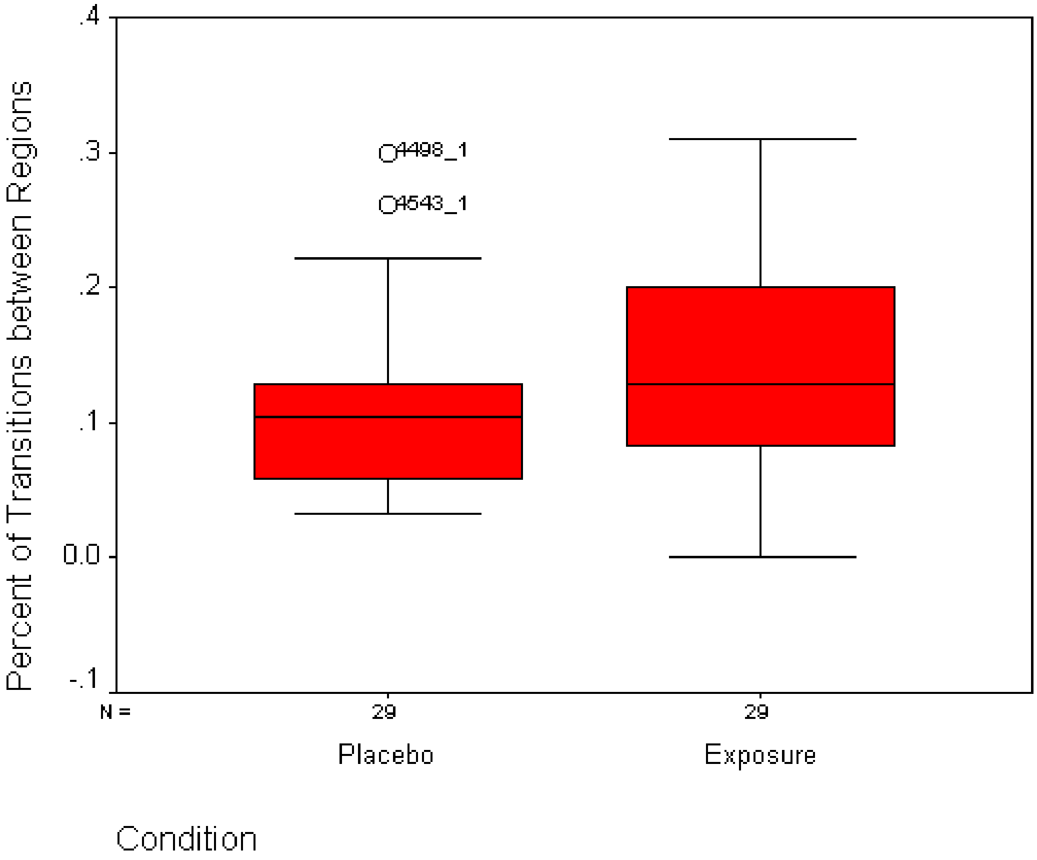

Figure 5.6: Condition as related to discourse organization in the dog picture

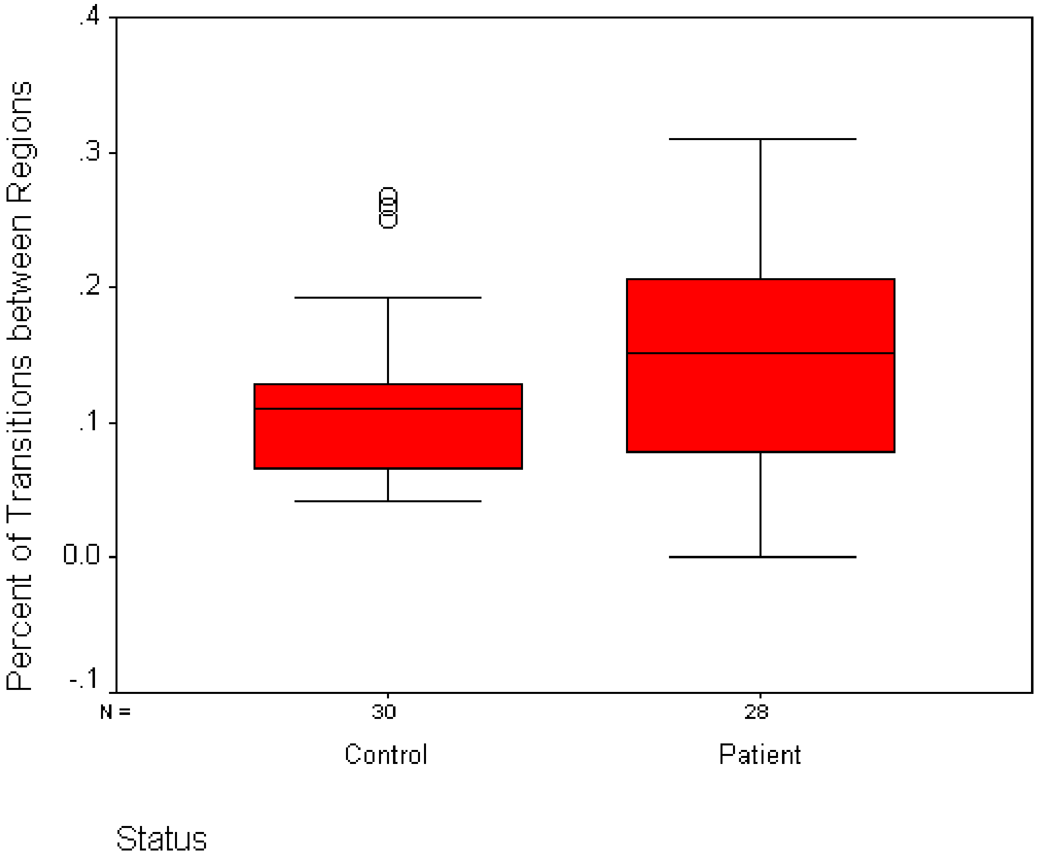

Figure 5.7: Status as related to discourse organization in the dog picture

Chapter 6

Discussion and Conclusions

6.1 Discussion of Findings

In this study, results have suggested that the Salient Items Test might be a useful measurement for diagnosing certain forms of schizophrenia. In the horse picture, the total number of salient items mentioned was a useful indicator; in the dog picture, the average position of these items was more indicative. Results also confirm earlier studies that disorganized discourse, as can be measured through total number of transitions, correlates with schizophrenia.

Different pictures, however, yield different results. One might argue that the differing numbers of salient items in the pictures or the differing organization of the pictures themselves has contributed to this effect. Perhaps the lower frequencies of salient items in the horse picture corpus translate into individual descriptions that are more likely to leave out items. Perhaps the clearer organization of the dog picture means a more predictable order and timing in naming salient items, making it easier to discriminate between normal and patient groups. These results, though different, seem to suggest that salient items might be a useful measurement for distinguishing between schizophrenic patients and healthy controls. However, given the differing results, a larger follow up study would be necessary to confirm any results.

Perhaps even more important than these immediate findings as related to schizophrenia are those related to the actual methodology. Selection of picture is clearly an important variable in picture descriptions; one is not always as good as the next. Different pictures bring out different linguistic features, and make studying some features more or less difficult.

This study found a higher level of odd types in the horse picture than in the dog picture, and correspondingly fewer salient items and more variability in discourse organization. A high level of odd types indicates a less centralized lexicon; pictures that create this effect are more variable in interpretation. Pictures with no clear spatial organization create a greater variety of discourse organizations, making organization less easy to study. Clearly what picture a person describes has a big impact on his description.

For future computerized studies of discourse organization in picture descriptions, results from this study suggest that a picture modeled off the dog picture might be ideal. Key features in such a picture might include two clear discreet regions, each containing clear, discreet salient items. By confining salient items and interaction between these items to two distinct regions, and avoiding interaction of items between the regions, it becomes potentially possible to measure transitions between regions with relative ease. As this study suggests that an elevated level of inter-region transitions correlates with schizophrenia, such a figure might serve as a useful index for detecting one of the symptoms of schizophrenia, disorganized discourse.

With computer analysis, reducing variability in responses is especially crucial. By constraining the population of normal responses, it is easier to identify remarkable ones. The clearer the picture, the more it should lend itself to salient items and organizational studies. Though the TAT pictures and the “cookie theft” picture are more canonical, therefore having a well-documented canon of normal responses, they may not be ideal for every type of linguistic study. If different pictures can bring out different linguistic features, as seen in the pictures designed for this study, more attention should be paid to designing pictures with the specific test hypothesis in mind. Better pictures could mean better results.

Bibliography

Covington, Michael. (2004) Computer Implemented Vocabulary-based Test for a Class of Language Disorder. *US Patent Application 20040054266.*

Croisile, Bernard; Ska, B.; Brabant, M. J.; Duchene, A.; Lepage, Y.; Aimard, G.; and Trillet, M. (1996) Comparative Study of Oral and Written Picture Description in Patients with Alzheimer's Disease. *Brain and Language*, 53: 1-19.

Docherty, Nancy. (2005) Cognitive Impairments and Disordered Speech in Schizophrenia: Thought Disorder, Disorganization, and Communication Failure Perspectives. *Journal of Abnormal Psychology*, 114: 269-278.

Goodglass, Harold, and Kaplan, Edith. (1984) *The Assessment of Aphasia and Related Disorders.* Lea and Febiger.

Gernsbacher, Morton Ann; Tallent, Kathleen; and Bollinger, Caroline. (1999) Disordered Discourse in Schizophrenia Described by the Structure Building Framework. *Discourse Studies*, 1(3): 355-372.

Goren, Anat; Tucker, Gary; and Ginsberg, Gary. (1996) Language Dysfunction in Schizophrenia. *European Journal of Disorders of Communication*, 31: 153-170.

Lee, Jay, and Kretzschmar, William. (1993) Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Information Systems*, 7(6): 541-560.

Liddle, P. F. (1987) The Symptoms of Chronic Schizophrenia: A Re-examination of the Positive-Negative Dichotomy. *British Journal of Psychiatry*, 151: 145-151.

McKenna, Peter, and Oh, Tomasina. (2005) *Schizophrenic Speech.* Cambridge University Press.

Morgan, Wesley. (1995) Origin and History of the Thematic Apperception Test Images. *Journal of Personality Assessment*, 65(2): 237-254.

Oh, T. M.; McCarthy, R. A.; and McKenna, P. J. (2002) Is There a Schizophasia? A Study Applying the Single Case Approach to Formal Thought Disorder in Schizophrenia. *Neurocase*, 8: 233-244.

Osselton, N., and Hempelman, R. (2003) *The New Routledge Dutch Dictionary.* London: Routledge.

Parker, Philip. (2005) *Webster's Online Dictionary.* <www.websters-online-dictionary.org>.

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Smith, Rebecca; Heuerman, Maranda; Wilson, Brenda; and Proctor, Adele. (2003) Analysis of Normal Discourse Patterns. *Brain and Cognition*, 53(2): 368-371.

Appendix A

Case Study Transcripts

A.1 Description 4483_2

57001
Ok, ehh. Ik zie een berglandschap
Ok, ehh. I see a mountainous landscape

60001
voor me. Ehh
before me. Ehh

63001
vooraan op de afbeelding
in the front of the picture

66001
eh ziet ge een ruiter op zijn paard,
eh you can see a rider on his horse,

69001
op een eh, bergwand staan. En eh
on a eh, mountain side. And eh

72001
die overziet een vallei. In die vallei
that overlooks a valley. In that valley

75001
eh, is een spoorweg en daar rijdt momenteel een
eh, is a railway and at the moment there's a

78001
trein op. Ehh..
train riding on it. Ehh..

81001
het is een, zonnig, de zon schijnt. Er zijn

it is a, sunny, the sun is shining. There are

84001
geen wolken aan de lucht en er is ook een vogeltje aan de lucht,
no clouds in the sky and there is a little bird on the sky as well,

87001
in de lucht. En..
in the sky. And..

90001
over de ruiter ehh.. hij
about the rider ehh.. he

93001
steekt zijn hand op. Alsof hij ergens
is raising his hand. As if he is

96001
naar aan het wuiven is. Naar het waaien of
waving at something. Waving at or

99001
hoe noem je dat in Nederland. Ehh..
how do you call that in the Netherlands. Ehh..

102001
Hij
He

105001
zit op zijn paard en zijn paard heeft z'n rechter
is sitting on his horse and his horse has its right

108001
voorpoot ehh
forefoot ehh

111001
lichtjes omhoog geheven. Ehh
lifted up a little. Ehh

114001
de ruiter houdt de teugels strak want
the rider is holding the reins tight because

117001

het paard trekt een beetje naar achter met zijn hoofd.
the horse is pulling its head back a little.

120001
Ehh,
Ehh,

123001
de trein in de vallei, ehh
the train in the valley, ehh

126001
ziet er uit als een eh TGV.
looks like a eh TGV.

129001
Een vrij moderne trein. Ik vind het een rare
A pretty modern train. I think it's a strange

132001
combinatie trouwens. Zo'n ruiter met zo'n
combination by the way. Such a rider with such a

135001
trein, moderne trein. Ik zou zo'n oude
train, modern train. I would

138001
western trein verwachten eerlijk gezegd. Ehh
expect an old western train to be honest. Ehh

141001
het is een
it is a

144001
ehh, in de vallei is het precies een woestijn
ehh, in the valley it looks like a desert

147001
landschap. Er staat een cactus en wat
landscape. There is a cactus and some

150001
verdorde struikjes. Ehh
small withered bushes. Ehh

153001
Ehh
Ehh

156001
de vogel links in de
the bird left in the

159001
boven in de lucht is ehh
up in the sky is ehh

162001
zwart heeft een wit hoofd precies.
black has and a white head exactly.

165001
Ehh het eerste wat bij mij opkomt
Ehh the first thing that comes to my mind

168001
was een zwaluw alhoewel dat, dat ook
was a swallow although that, that

171001
eventueel iets anders zou kunnen zijn.
could possibly be something else as well.

174001
Ehh de ruiter, om er op terug te komen,
Ehh the rider, to get back to that,

177001
heeft ehh lange zwarte rij-
is wearing ehh long black riding

180001
laarzen aan. Ehh een lange witte mantel.
boots. Ehh a long white coat.

183001
Ehh
Ehh

186001
wat klassiek kapsel.
a bit of a classic hairstyle.

189001

192001

195001
Ehh, op het eerste gezicht vraagt ge meteen af
Ehh, at first glance you immediately wonder

198001
waar hij naar aan het wuiven
what he’s waving at

201001
is. En, een mogelijk scenario zou
. And, a possible scenario would

204001
zijn dat hij teken geeft aan zijn
be that he’s giving a sign to his

207001
kameraden. Misschien dat hij een leider is van een
comrades. Maybe he’s a leader of a

210001
roversbende, om eh om over
gang of robbers, to eh about to go

213001
te gaan de trein overvallen. Omdat die net aangekomen is.
to raid the train. Because it has just arrived.

216001
En dat ze al een tijdje hebben zitten wachten.
And that they have been waiting for some time.

219001
Ehh..
Ehh..

222001

225001

A.2 Description 4498_2

66001
Het gaat zich dus over een
It’s obviously about a

69001
landschap.
landscape.

72001
Waar een trein
Where a train

75001
doorheen rijdt.
is riding through.

78001

81001
Het is vandaag een zomerse dag.
Today is a summery day.

84001

87001

90001
En iemand is met een
And somebody’s

93001
paard er op uit.
riding on a horse.

96001
Om de natuur te verkennen.
Exploring the scenery.

99001

102001

105001
Hij wijst naar de trein. Misschien dat hij dat hinderlijk vindt dat daar de trein
He’s pointing at the train. Maybe he’s annoyed by the train

108001
rijdt.
riding there.

111001

114001

117001

120001

123001
(proefleider: is er nog meer dat je zou kunnen
(experiment leader: is there anything else that you could

126001
vertellen over het plaatje?)
say about the picture?)

129001

132001
Ik denk dat het paard ook schrikt van iets.
I think the horse is also startled by something.

135001
Of het van de trein is of
It’s either by the train or

138001
van de afgrond is.
it’s by the precipice.

141001
Weet
I don’t

144001
ik niet. Maar ik
know. But I

147001
denk, zoals op het plaatje te zien is dat het paard van
think, as can be seen from the picture, that the horse is

150001
iets schrikt.
startled by something.

153001

156001

159001

162001
Het is een droog gebied.
It’s a dry region.

165001
Er groeien cactussen.

There are cacti growin.

168001

171001

174001

177001

180001

183001

186001
(proefleider: zie je nog meer dingen in het plaatje?)
(experiment leader: can you see anything else in the picture?)

189001

192001

195001
Ja een landschap
Yes a landscape

198001
met kale vlaktes.
with barren plains.

201001

204001

A.3 Description 4500_1

90001
(onverstaanbaar)
(unintelligible)

93001

96001
Een bus ervoor..
A bus in front of it..

99001
Bus ervoor
Bus in front of it

102001

105001
(onverstaanbaar) op de grond
(unintelligible) on the ground

108001
een hond erbij. En een vrouw
a dog there. And a woman

111001
die de hond uitlaat. En eh..
that is walking the dog. And eh..

114001
paraplu...
umbrella...

117001

120001

A.4 Description 4500_2

90001
Ik zie een..
I see a..

93001
paard met een man erop.
horse with a man on it.

96001
Een zadel,
A saddle,

99001
halster.
halter.

102001
Een trein zie ik.
A train I see.

105001
Een cactus (onverstaanbaar)
A cactus (unintelligible)

108001

111001
Bergen.
Mountains.

114001

117001
Hoge bergen.
High mountains.

120001

123001
Een zon.
A sun.

126001
Een vogeltje nog.
A bird as well.

129001

132001

A.5 Description 4458_1

54001
Ehh, ik zie een zwart wit plaatje.
Ehh, I can see a black and white picture.

57001
Op het plaatje van links naar rechts zie ik een
On the picture from left to right I can see a

60001
ehh, bloem, zie ik bloemen, tulpen.
ehh, flower, I can see flowers, tulips.

63001
(onverstaanbaar) Ehh, met ehh, kelkbladeren.
(unintelligible) Ehh, with ehh, chalice-like petals.

66001
Ehh, aan die bloemen ruikt
Ehh, a dog is smelling the flowers

69001
een hond, het lijkt op een golden retriever.
, it looks like a Golden Retriever.

72001
Hij heeft een halsband die
He’s got a collar that

75001

ook om zijn midden gaat. Ehh,
goes around its waist as well. Ehh,

78001
Ehhm, hij heeft een
Ehm, he's got a

81001
zwarte neus, eh...
black nose, eh...

84001
Eh, de hond wordt vastgehouden door
Eh, the dog is being held by

87001
een eh, door een vrouw met kort haar.
a eh, by a woman with short hair.

90001
Ehh, ze draagt een blouse met strepen
Ehh, she's wearing a blouse with stripes

93001
en een rok, die toch ongeveer tot aan haar
and a skirt, approximately to her

96001
knien komt. Ehh, ze heeft
knees. Ehh, she's wearing

99001
ja, soort klompen aan lijkt het.
well, some sort of clogs, it seems.

102001
Ehh, ze heeft een paraplu vast.
Ehh, she's holding an umbrella.

105001
En er zijn, het lijkt er ook op
And there are, it looks like

108001
alsof het regent, omdat er zijn eh,
it's raining, because there are eh,

111001
strepen door het eh scherm heen. Ehh..
stripes through the eh screen. Ehh...

114001
Zij loopt op een paadje.
She’s walking on a small path.

117001
Het lijkt een landweggetje. En op dat paadje ligt een
It looks like a small country road. And on that small path is a

120001
plas. Ehhm,
puddle. Ehm...

123001
achter de hond eh, staat een grote kerk.
behind the dog eh, is a big church.

126001
Ehh, waar op twee klokken zijn te zien.
Ehh, where two clocks can be seen.

129001
Twee ramen en een spits
Two windows and a pointed

132001
dak, een spitse toren. Eh,
roof, a pointed tower. Eh,

135001
voor die kerk staat een eh,
in front of that church is a eh,

138001
een bus, met zes
a bus, with six

141001
of ze-, zeven ramen. Eh
or se-, seven windows. Eh

144001
met eh lampen.
with eh, lamps.

147001
En wielen, ehh..
And wheels, ehh..

150001
Het geheel is ehh,
The whole (picture) is ehh...

153001
heeft perspectief, ehh
is in perspective, ehh

156001
de horizon ligt ongeveer in het midden.
the horizon is roughly in the middle.

159001
Eh, op de horizon
Eh, on the horizon

162001
zie je heel vaag eh..
you can vaguely see eh...

165001
bergen of huizen, ja niet bergen maar
mountains or houses, well not mountains but

168001
struiken en huizen op de achtergrond.
bushes and houses in the background.

171001
Ehhm,
Ehmm..

174001
kijken, de kerk bestaat uit
let’s see, the church consists of

177001
ongeveer, ja
approximately, well

180001
wa-, wat ik nu zie drie blokken en n toren die in het midden

wha-, what I can see now, three blocks and one tower in the middle

183001
staat. Ehh..
. Ehh...

186001
de bus staat aan de linkerkant ervan.
the bus is on the left side of this.

189001
Ehm,
Ehm,

192001

195001
Het gras dat, zitten geen duidelijke
The grass, there are no clear

198001
ehm details in. Dus je
ehm details in it. So you

201001
ziet niet de grassprieten alleen maar heel klein
can’t see the blades of grass, just a little

204001
beetje bij eh, bij de tulpen.
bit near eh, near the tulips.

207001
Ehm,
Ehm,

210001
de vrouw die heeft een eh, tevreden blik op haar
the woman has a eh, contented look on her

213001
gezicht. Eh..
face. Eh...

216001

D'r haar zit achter d'r oor. Ehh,
Her hair is behind her ear. Eh..

219001
het lijkt alsog ze d'r o, 'd'r ogen dicht heeft.
it looks as if she has her e-, her eyes closed.

222001
Ehm,
Ehm...

225001
ze houdt de riem, of de lijn van de hond vast met
she's holding the lead, or the dog's leash with

228001
twee handen. Waarbij ze de rechter
two hands. And her right

231001
hand eh, de riem om haar pink laat gaan.
hand, eh she's holding the lead around her pinky.

234001
Ehm,
Ehm...

237001

240001
Ja en de bus die, daar zie je nu twee
Well and the bus, you can now see two...

243001

246001

A.6 Description 4619_1

60001
Ik zie een mevrouw. Die is haar hond uit aan het laten.
I can see a lady. She's walking her dog.

63001
Voor die hond zie ik vijf tulpen.
In front of the dog I can see five tulips.

66001
De hond is een blindengeleidehond waarschijnlijk,
The dog is probably a guide dog for the blind,

69001
want hij heeft een speciaal tuig om. Het regent op
because he's wearing a special halter. It's raining on

72001
die dag. Die mevrouw heeft een paraplu op.
this day. The woman has an umbrella up.

75001
De vrouw heeft een eh, jurkje
The woman is wearing a eh, small dress

78001
aan en een soort hempje.
and a sort of vest.

81001
Op de achtergrond is een eh, bus
In the background there's a eh, bus

84001
en achter de bus een kerk. Op de klok
and behind the bus a church. The church clock

87001
van de kerk is het tien voor n.
shows ten to one.

90001
Het soort hond wat ik zie is
The kind of dog that I see is

93001
eh denk ik een golden retriever.
eh a Golden Retriever I think.

96001

99001
En op de bus
And on the bus

102001
staat niks welke, van welke firma dat ze is.
it doesn't say which, which company it's from.

105001
Ik kan de mevrouw beter proberen te
I'd better try to describe the lady

108001
beschrijven.
.

111001
Ze houdt met haar linkerhand de paraplu vast, met haar
With her left hand she's holding her umbrella, with her

114001
rechterhand houdt ze de lijn vast van de hond.
right hand she's holding the dog's lead.

117001
De hond is aan het snuffelen aan
The dog's sniffing

120001
de bloemen.
the flowers.

123001
Op de achtergrond zie ik nog vaag bergen.
In the background I can vaguely see mountains as well.

126001
Ik weet niet of dat bergen zijn maar ik zie lijn-
I don't know if those are mountains but I can see

129001
tjes. En de vrouw loopt op een pad.
lines. And the woman's walking on a path.

132001

In het midden van het pad is waarschijnlijk
In the middle of the path there's

135001
een plas.
a puddle probably.

138001

141001

144001

147001

150001

153001
(proefleider: is er nog meer wat je zou kunnen vertellen?)
(experiment leader: is there anything else that you could tell?)

156001
Ja hoe de kerk er uitziet. Eh..
Yes what the church looks like. Eh...

159001
of dat de vrouw een gestreept eh,
or that the woman is wearing a striped eh,

162001
eh jasje cq hemd aan heeft.
eh jacket or shirt.

165001
Eh, de
Eh, the

168001
vrouw heeft d’r haren de scheiding in het midden.
woman’s hair has a parting in the middle.

171001
Ze, het lijkt wel of dat ze een beetje
She, it looks like she’s smiling

174001
lacht. Een kuiltje in,
somewhat. A dimple in,

177001
in haar wang. Ze heeft
in her cheek. She’s got

180001
een ja, het lijkt veel, het zijn geen klompen maar het lijken
a yes, it looks a lot like, those are not clogs but they look like

183001
wel op klompen wat ze aan heeft.
clogs she’s wearing.

186001

189001
In de kerk zie ik eh ja,
In the church I can see eh yes,

192001
aan mijn kant twee ramen.
on my side two windows.

195001
Grote ramen, n langwerpig,
Big windows, one oblong,

198001
En de
And the

201001
n wat minder, is hetzelfde langwerpig ja. Het is nou
one is somewhat less, is oblong as well yes. It’s

204001
tien voor n, twaalf minuten voor n op de klok.
ten to one now, twelve minutes to one on the clock.

207001
In de bus kan ik alleen
In the bus I can only

210001
het stuur zien eh.. Ik zie
see the wheel eh... I can see

213001
daar een bus staan met zes ruiten, eh..
a bus standing there with six windows, eh...

216001
Dat was het.
That's all.

219001

222001
Waarschijnlijk is die vrouw, is die vrouw blind.
The woman is, the woman is probably blind.

225001

228001
Of slechtziend.
Or visually impaired.

231001

234001

A.7 Description 4470_1

78001

Een vrouw
A woman

81001
wandelt op
is walking on

84001
een voetpaadje met haar
a small footpath with her

87001
hond. Het regent, zodat ze
dog. It's raining, so she

90001
een paraplu boven haar hoofd
has put up an umbrella over her head.

93001
draagt.

96001
De hond ruikt aan wat tulpen,
The dog smells some tulips,

99001
die naast het paadje
that are growing next to the small path.

102001
groeien.

105001
Op de achtergrond
In the background

108001
zie je een uh kerk staan,
you can see a [uh] church,

111001
waar de bus langs komt.
which the bus is passing.

114001

117001
De kerk wijst een uur of uh
The church shows that it is about [uh]

120001
tien voor twee aan.
ten to two.

123001

126001
Het is erg licht uh,
It's very bright [uh],

129001
de zon schijnt
the sun is obviously shining

132001
kennelijk wel.

135001
De vrouw is niet extra dik
The woman isn't wearing particularly thick

138001
of warm aangekleed,
or warm clothing,

141001
maar het regent.
but it's raining.

144001

147001

150001

153001

156001
De hond staat stil
The dog stands still

159001
en ruikt aan
and smells

162001
de tulpen. De vrouw uh
the tulips. The woman [uh]

165001
de vrouw kijkt naar haar hond.
the woman watches her dog.

168001

171001

174001

177001
Ze zal wel doorlopen.
She'll probably walk on.

180001
en
and

183001
hopelijk voor uhh hopelijk re gaat het niet te hard regenen.

hopefully for [uh] hopefully it won’t rain too hard.

186001

189001

192001
Misschien moet de vrouw
Maybe the woman is

195001
wel uh naar de bus
on her way [uh] to the bus

198001
en ziet ze dat de bus stopt,
and she sees the bus stopping,

201001
zodat ze daar naar toe kan lopen en de bus
so she can walk over there and [uh] take the bus

204001
uh kan nemen

207001
naar   ergens anders.
to   somewhere else.

210001

213001

216001

219001

Moet ik blijven vertellen?
Should I continue?

222001

225001
Misschien moet gaat de vrouw gewoon langs de bus
Maybe the woman must... is simply going to walk past the bus

228001
naar de kerk.
to the church.

231001
Ik weet niet of een hond de kerk in
I don’t know if a dog is allowed inside a church.

234001
binnen mag.

237001

240001

[laughs]

243001

246001

Appendix B

Programs

B.1 Odd Types Program

```
def punctuation(character):
    "Checks if a character matches punctuation list, returns value."
    punctuation = [’.’,’,’,’:’,’;’,’!’,’?’,’\n’]
    #fptr = open(fname)
    #a = fptr.read()
    if character in punctuation:
        return 1
    else:
        return 0

def homogenizer(fname):
    "Removes punctuation and capitalization
         from a string."
    #fptr = open(fname)
    #a = fptr.read()
    new_string = ’’
    for character in fname:
        if punctuation(character) != 1:
            new_string = new_string + character
        elif punctuation(character) == 1:
            new_string = new_string + " "
    #fptr.close()
    file_list = new_string.lower().split()
    import Stemmer
    stemmed = Stemmer.Stemmer(’dutch’)
    cleaned = stemmed.stem(file_list)
    return cleaned

def type_counter(list):
    "Returns simple type list."
    list.sort()
    y = []
    frequency = []
    #subtracts 1 from total list length since the
    #1st postion is 0 to get
    #final postion
    prev = ""
    final_position = len(list) - 1
    #total number of words in list
    total = len(list)
    if total != 0:
        final_word = list[final_position]
        frequency.append(final_word)


    for token in list:
        if token != prev:
            frequency.append(prev)
            prev = token
        elif token == prev:
            prev = token
    frequency.sort()
    return frequency

def type_number(list):
    "Returns number of types."
    list.sort()
    y = []
    frequency = []
    #subtracts 1 from total list length since the
    #1st postion is 0 to get
    #final postion
    prev = ""
    final_position = len(list) - 1
    #total number of words in list
    total_length = len(list)
    if total_length != 0:
        final_word = list[final_position]
        frequency.append(final_word)

    for token in list:
        if token != prev:
            frequency.append(prev)
            prev = token
        elif token == prev:
            prev = token

    return len(frequency) + .0
```

```
def compare_uncommon(fname, norm):
    "Compares two files by type, returns
         uncommon words."
    fname_list = type_counter(homogenizer(fname))
    norm_list = type_counter(homogenizer(norm))
    really_odd_list = []

    for token in fname_list:
        if token not in norm_list:
            really_odd_list.append(token)

    position_list = []
    total_tokens = len(homogenizer(fname)) + .0
    for type in really_odd_list:
       total_type_tokens = homogenizer(fname).count(type)
       first_position = homogenizer(fname).index(type) + 1
       first_position_percent = first_position/total_tokens
       position_list.append(first_position_percent)


    ratio1 = len(really_odd_list)/total_tokens
    ratio2 = len(really_odd_list)/type_number(homogenizer(fname))

    sum = 0
    total_items = 0

    if len(position_list) >= 1:
        for number in position_list:
            sum = sum + number
            total_items = total_items + 1.0
        average_position = sum/total_items
    else:
        average_position = 0

    print str(len(really_odd_list))+’,’+str(total_tokens)+’,’
        +str(type_number(homogenizer(fname)))+
        ’,’+str(ratio1)+’,’+str(ratio2)+’,’
        +str(average_position)
```

B.2 Salient Items Program

```
def punctuation(character):
    "Checks if a character matches punctuation list
         and returns value."
    punctuation = [’.’,’,’,’:’,’;’,’!’,’?’,’\n’]
    #fptr = open(fname)
    #a = fptr.read()
    if character in punctuation:
        return 1
    else:
        return 0

def homogenizer(fname):
    "Removes punctuation and capitalization
         from a string."
    fptr = open(fname)
    file = fptr.read()
    new_string = ’’
    for character in file:
        if punctuation(character) != 1:
            new_string = new_string + character
        elif punctuation(character) == 1:
            new_string = new_string + " "
    #fptr.close()
    b = new_string.lower().split()
    import Stemmer
    stemmed = Stemmer.Stemmer(’dutch’)
    cleaned = stemmed.stem(b)
    return cleaned

def salient_dog(fname):
    "Returns information about salient items."
    dog = [’hond’, ’retriever’, ’gemengd’, ’hybridisch’,
             ’blindengeleidehond’,
           ’hondj’, ’retriever-acht’]
    church = [’kerk’, ’kerkj’,’kathedral’, ’dom’, ’kapel’,
                ’muziekkapel’, ’kerkgebouw’,
              ’staatskerk’, ’kerkgenootschap’, ’gebouw’,
               ’gebouwtjes’, ’bijgebouwtjes’,
              ’bouwwerk’, ’pand’]
    bus = [’bus’, ’autobus’, ’stadsbus’, ’tourbus’,
             ’personenvervoersbus’, ’schoolbus’,
           ’lijnbus’,’streekbuss’, ’streekbus’, ’buss’,
           ’reisbus’,’toeristenbus’,’stadslijn’]
    woman = [’vrouw’, ’werkster’, ’dienstmeid’, ’meid’,
             ’maitres’, ’dam’, ’wijfj’,
             ’vrouwtj’,’mevrouw’,’vrouwtjes’]
    umbrella = [’paraplu’, ’bescherm’,’parasol’]
    flowers = [’bloem’, ’tulp’, ’tulpjes’, ’bloemetjes’,
                 ’bloesem’,’bloemetj’]
    clogs = [’klomp’, ’klompjes’, ’klompacht’,’schoen’,
             ’hoefijzer’, ’remschoen’]
```

```
leash = [’riem’, ’lijn’, ’hondenlijn’, ’wandelriem’,
 ’krag’, ’boord’,
         ’halsboord’,’halsband’,’hondenriem’,
         ’aangelijnd’,’lijntjes’,’leibandj’,
         ’leiband’,’geleideband’, ’geleidebandj’]
rain = [’regent’, ’regen’, ’reg’, ’regendruppel’,
         ’regen-’, ’regenacht’]
puddle = [’plas’, ’plasj’, ’plass’, ’modderpoel’,
          ’poel’, ’regenplas’,’waterplas’]
clock = [’klok’, ’uurwerk’,’kerkklok’]
coat = [’overjas’, ’jas’, ’mantel’, ’jasj’,
         ’rok’, ’rokj’, ’colbert’, ’onhulsel’,
          ’bekled’, ’hul’]
grass = [’gras’, ’gazon’, ’grasveld’, ’batist’,
         ’erf’]
path = [’weg’, ’pad’, ’paadj’, ’ban’, ’rout’,
         ’voetpad’, ’voetpaadj’,
        ’voetgangerspad’,’landweg’,
        ’landweggetj’,’wandelpad’,’wandelpaadj’,
        ’weggetj’,’zandweggetj’,’zandweg’]
walk = [’loopt’,’wandel’, ’wandelt’, ’marcher’,
        ’tippel’, ’lop’]
tower = [’tor’, ’kerktor’]




item_list = [dog, church, bus, woman, umbrella,
             flowers, clogs, leash, rain, puddle,
             clock, coat, grass, path, walk, tower]

total_number_of_salient_items = 0
weighted_salient_items = 0
total_salient_items_tokens = 0
salient_items_found = []
total_first_positions = []
match_word_list = []
all_items_first_positions = []
all_salient_items_tokens = []

for item in item_list:
    item_first_positions = []
    for word in item:
        if word in homogenizer(fname):
            all_salient_items_tokens.append(word)
            word_total = homogenizer(fname).count(word)
            total_salient_items_tokens = word_total +
                                 total_salient_items_tokens

            word_first_position = homogenizer(fname).index(word)
            item_first_positions.append(word_first_position)
        #else:
            #for token in homogenizer(fname):
                #if word in token and token not in item:
                    #print token + " " + word


    if len(item_first_positions) > 1:
        item_first_positions.sort()
        the_first_position = item_first_positions[0]
        total_first_positions.append(the_first_position)
        match_word = homogenizer(fname)[the_first_position]
        match_word_list.append(match_word)
        total_number_of_salient_items =
                        total_number_of_salient_items + 1
    elif len(item_first_positions) == 1:
        the_first_position = item_first_positions[0]
        total_first_positions.append(the_first_position)
        match_word = homogenizer(fname)[the_first_position]
        match_word_list.append(match_word)
        total_number_of_salient_items =
                        total_number_of_salient_items + 1


total_tokens = len(homogenizer(fname)) + .0
position_sum = 0
total_items = 0
sum = 0
for position in total_first_positions:
    sum = sum + position
    total_items = total_items + 1.0
average_first_position = sum/total_items
salient_items_found = []
weighted_items = 0

dog_position = "-"
church_position = "-"
bus_position = "-"
woman_position = "-"
umbrella_position = "-"
flower_position = "-"
clogs_position = "-"
leash_position = "-"
rain_position = "-"
```

```
puddle_position = "-"
clock_position = "-"
coat_position = "-"
grass_position = "-"
path_position = "-"
walk_position = "-"
tower_position = "-"

for word in match_word_list:
    if word in dog:
        salient_items_found.append('dog')
        dog_position = homogenizer(fname).index(word)
    elif word in church:
        salient_items_found.append('church')
        church_position = homogenizer(fname).index(word)
    elif word in bus:
        salient_items_found.append('bus')
        bus_position = homogenizer(fname).index(word)
    elif word in woman:
        salient_items_found.append('woman')
        woman_position = homogenizer(fname).index(word)
    elif word in umbrella:
        salient_items_found.append('umbrella')
        umbrella_position = homogenizer(fname).index(word)
    elif word in flowers:
        salient_items_found.append('flowers')
        flower_position = homogenizer(fname).index(word)
    elif word in clogs:
        salient_items_found.append('clogs')
        clogs_position = homogenizer(fname).index(word)
    elif word in leash:
        salient_items_found.append('leash')
        leash_position = homogenizer(fname).index(word)
    elif word in rain:
        salient_items_found.append('rain')
        rain_position = homogenizer(fname).index(word)
    elif word in puddle:
        salient_items_found.append('puddle')
        puddle_position = homogenizer(fname).index(word)
    elif word in clock:
        salient_items_found.append('clock')
        clock_position = homogenizer(fname).index(word)
    elif word in coat:
        salient_items_found.append('coat')
        coat_position = homogenizer(fname).index(word)
    elif word in grass:
        salient_items_found.append('grass')
        grass_position = homogenizer(fname).index(word)
    elif word in path:
        salient_items_found.append(’path’)
        path_position = homogenizer(fname).index(word)
    elif word in walk:
        salient_items_found.append(’walk’)
        walk_position = homogenizer(fname).index(word)
    elif word in tower:
        salient_items_found.append(’tower’)
        tower_position = homogenizer(fname).index(word)



if ’pcd’ in fname:
    print ’p,c,d,’+str(fname)+’,’+
    str(total_number_of_salient_items)
    +’,’+
    str(total_tokens)+’,’+str(total_salient_items_tokens)
    +’,’
    +str(average_first_position)+
    ’,’+str(average_first_position/total_tokens)+’,’
    +str(dog_position)+’,’
    +str(church_position)+’,’+str(bus_position)+’,’+
    str(woman_position)+
    ’,’+str(umbrella_position)+
    ’,’+str(flower_position)+’,’+str(clogs_position)+’,’
    +str(leash_position)
    +’,’+str(rain_position)+’,’
    +str(puddle_position)+’,’+str(clock_position)+’,’
    +str(coat_position)
    +’,’+str(grass_position)+’,’
    +str(path_position)+’,’+str(walk_position)+’,’+
    str(tower_position)

elif ’ecd’ in fname:
    print ’e,c,d,’+str(fname)+’,’+
    str(total_number_of_salient_items)
    +’,’+ str(total_tokens)+
    ’,’+str(total_salient_items_tokens)+’,’+
    str(average_first_position)+
    ’,’+
    str(average_first_position/total_tokens)+
    ’,’+str(dog_position)
    +’,’+str(church_position)+
    ’,’+str(bus_position)+’,’+str(woman_position)+’,’
    +str(umbrella_position)+’,’
    +str(flower_position)+’,’+str(clogs_position)+
    ’,’+str(leash_position)+
    ’,’+str(rain_position)+
    ’,’+str(puddle_position)+’,’+
    str(clock_position)+’,’
    +str(coat_position)+’,’
    +str(grass_position)+’,’+str(path_position)+’,’+
    str(walk_position)+
    ’,’+str(tower_position)

elif ’ppd’ in fname:
    print ’p,p,d,’+str(fname)+’,’+
    str(total_number_of_salient_items)+’,’+
    str(total_tokens)+’,’+str(total_salient_items_tokens)
    +’,’+
    str(average_first_position)+’,’
    +str(average_first_position/total_tokens)+’,’
    +str(dog_position)+’,’
    +str(church_position)+
    ’,’+str(bus_position)+’,’+str(woman_position)+’,’
    +str(umbrella_position)+’,’
    +str(flower_position)+’,’+str(clogs_position)+’,’
    +str(leash_position)
    +’,’+str(rain_position)
    +’,’+str(puddle_position)+’,’+str(clock_position)+
    ’,’+str(coat_position)
    +’,’+str(grass_position)
    +’,’+str(path_position)+’,’+str(walk_position)+
    ’,’+str(tower_position)

elif ’epd’ in fname:
    print ’e,p,d,’+str(fname)+’,’+
    str(total_number_of_salient_items)+’,’
    + str(total_tokens)+’,’
    +str(total_salient_items_tokens)+’,’+
    str(average_first_position)+’,’+
    str(average_first_position/total_tokens)+’,’
    +str(dog_position)+’,’+
    str(church_position)+
    ’,’+str(bus_position)+’,’+str(woman_position)+’,’
    +str(umbrella_position)+
    ’,’+str(flower_position)+
    ’,’+str(clogs_position)+’,’+str(leash_position)+
    ’,’+str(rain_position)+
    ’,’+str(puddle_position)
    +’,’+str(clock_position)+’,’+str(coat_position)+’,’
    +str(grass_position)+
    ’,’+str(path_position)+’,’
```

```
        +str(walk_position)+’,’+str(tower_position)
```

B.3 Discourse Organization Program

```
def punctuation(character):
    "Checks if a character matches punctuation list
         and returns value."
    punctuation = [’.’,’,’,’:’,’;’,’!’,’?’,’\n’]
    #fptr = open(fname)
    #a = fptr.read()
    if character in punctuation:
        return 1
    else:
        return 0

def homogenizer(fname):
    "Removes punctuation and capitalization
         from a string."
    fptr = open(fname)
    file = fptr.read()
    new_string = ’’
    for character in file:
        if punctuation(character) != 1:
            new_string = new_string + character
        elif punctuation(character) == 1:
            new_string = new_string + " "
    fptr.close()
    word_list = new_string.lower().split()
    import Stemmer
    stemmed = Stemmer.Stemmer(’dutch’)
    cleaned = stemmed.stem(word_list)
    return cleaned


def type_counter(list):
    "Returns simple type list."
    list.sort()
    y = []
    frequency = []
    #subtracts 1 from total list length since
         the 1st postion is 0 to get
    #final postion
    prev = ""
    final_position = len(list) - 1
    #total number of words in list
    total = len(list)
    if total != 0:
        final_item = list[final_position]
        frequency.append(final_item)

    for token in list:
        if token != prev:
            frequency.append(prev)
            prev = token
        elif token == prev:
            prev = token
    frequency.sort()
    return frequency

def type_number(list):
    "Returns number of types."
    list.sort()
    y = []
    frequency = []
    #subtracts 1 from total list length since
         the 1st postion is 0 to get
    #final postion
    prev = ""
    final_position = len(list) - 1
    #total number of words in list
    total_length = len(list)
    if total_length != 0:
        final_item = list[final_position]
        frequency.append(final_item)

    for token in list:
        if token != prev:
            frequency.append(prev)
            prev = token
        elif token == prev:
            prev = token

    return len(frequency) + .0



def bigram_dog(fname):
    "Produces bigrams from the ordered salient items list."

    salient_items = ['hond', 'retriever', 'gemengd',
             ’hybridisch’, ’blindengeleidehond’,
        ’hondj’, ’retriever-acht’,
        ’kerk’, ’kerkj’,’kathedral’, ’dom’, ’kapel’,
        ’muziekkapel’, ’kerkgebouw’,
        ’staatskerk’, ’kerkgenootschap’, ’gebouw’,
        ’gebouwtjes’, ’bijgebouwtjes’,
        ’bouwwerk’, ’pand’,
        ’bus’, ’autobus’, ’stadsbus’, ’tourbus’,
        ’personenvervoersbus’, ’schoolbus’,
        ’lijnbus’, ’streekbuss’, ’streekbus’,’buss’,
        ’reisbus’,’toeristenbus’,’stadslijn’,
        ’vrouw’, ’werkster’, ’dienstmeid’, ’meid’,
        ’maitres’, ’dam’, ’wijfj’,
        ’vrouwtj’,’vrouwtjes’,’mevrouw’,
        ’paraplu’, ’bescherm’, ’parasol’,
        ’bloem’, ’tulp’, ’tulpjes’, ’bloemetjes’,
        ’bloemetj’, ’bloesem’,
        ’klomp’, ’klompjes’, ’klompacht’,’schoen’,
        ’hoefijzer’, ’remschoen’,
        ’riem’, ’lijn’, ’hondenlijn’, ’wandelriem’,
        ’krag’, ’boord’, ’leibandj’,
        ’geleideband’, ’geleidebandj’,’leiband’,
        ’halsboord’,’halsband’,’hondenriem’,
        ’aangelijnd’,’lijntjes’,
        ’regent’, ’regen’, ’reg’, ’regendruppel’,
        ’regen-’, ’regenacht’,
        ’plas’, ’plasj’, ’plass’, ’modderpoel’,
        ’poel’, ’regenplas’,’waterplas’,
        ’klok’, ’uurwerk’,’kerkklok’,
        ’overjas’, ’jas’, ’mantel’, ’jasj’,
        ’rok’, ’rokj’, ’colbert’, ’onhulsel’,
        ’bekled’, ’hul’,
        ’gras’, ’gazon’, ’grasveld’, ’batist’, ’erf’,
        ’weg’, ’pad’, ’paadj’, ’ban’, ’rout’,
        ’voetpad’, ’voetpaadj’,
        ’voetgangerspad’,’landweg’,
        ’landweggetj’,’wandelpad’,’wandelpaadj’,
        ’weggetj’,’zandweggetj’, ’zandweg’,
        ’loopt’,’wandel’, ’wandelt’, ’marcher’,
        ’tippel’, ’lop’,
        ’tor’, ’kerktor’]

    dog = [’hond’, ’retriever’, ’gemengd’,
         ’hybridisch’, ’blindengeleidehond’,
           ’hondj’, ’retriever-acht’]
    church = [’kerk’, ’kerkj’,’kathedral’, ’dom’,
         ’kapel’, ’muziekkapel’, ’kerkgebouw’,
              ’staatskerk’, ’kerkgenootschap’,
               ’gebouw’, ’gebouwtjes’, ’bijgebouwtjes’,
              ’bouwwerk’, ’pand’]
    bus = [’bus’, ’autobus’, ’stadsbus’, ’tourbus’,
         ’personenvervoersbus’, ’schoolbus’,
           ’lijnbus’,’streekbuss’, ’streekbus’,
            ’buss’,’reisbus’,
           ’toeristenbus’,’stadslijn’]
    woman = [’vrouw’, ’werkster’, ’dienstmeid’,
             ’meid’, ’maitres’, ’dam’, ’wijfj’,
             ’vrouwtj’,’mevrouw’,’vrouwtjes’]
    umbrella = [’paraplu’, ’bescherm’,’parasol’]
    flowers = [’bloem’, ’tulp’, ’tulpjes’,
             ’bloemetjes’, ’bloesem’,’bloemetj’]
    clogs = [’klomp’, ’klompjes’, ’klompacht’,
            ’schoen’, ’hoefijzer’, ’remschoen’]
    leash = [’riem’, ’lijn’, ’hondenlijn’,
             ’wandelriem’, ’krag’, ’boord’,
             ’halsboord’,’halsband’,’hondenriem’,
             ’aangelijnd’,’lijntjes’,’leibandj’,
             ’leiband’,’geleideband’, ’geleidebandj’]
    rain = [’regent’, ’regen’, ’reg’,
             ’regendruppel’, ’regen-’, ’regenacht’]
    puddle = [’plas’, ’plasj’, ’plass’, ’modderpoel’,
             ’poel’, ’regenplas’,’waterplas’]
    clock = [’klok’, ’uurwerk’,’kerkklok’]
    coat = [’overjas’, ’jas’, ’mantel’, ’jasj’,
             ’rok’, ’rokj’, ’colbert’,
             ’onhulsel’, ’bekled’, ’hul’]
    grass = [’gras’, ’gazon’, ’grasveld’,
            ’batist’, ’erf’]
    path = [’weg’, ’pad’, ’paadj’, ’ban’, ’rout’,
             ’voetpad’, ’voetpaadj’,
            ’voetgangerspad’,’landweg’,
            ’landweggetj’,’wandelpad’,’wandelpaadj’,
            ’weggetj’,’zandweggetj’,’zandweg’]
    walk = [’loopt’,’wandel’, ’wandelt’, ’marcher’,
             ’tippel’, ’lop’]
    tower = [’tor’, ’kerktor’]

    dutch_match_words = []
    salient_items_found = []
    coded_salient_items = []

    for token in homogenizer(fname):
        if token in salient_items:
            dutch_match_words.append(token)
```

```
for word in dutch_match_words:
    if word in dog:
        #salient_items_found.append('dog')
        coded_salient_items.append('1')
    elif word in church:
        #salient_items_found.append('church')
        coded_salient_items.append('2')
    elif word in bus:
        #salient_items_found.append('bus')
        coded_salient_items.append('2')
    elif word in woman:
        #salient_items_found.append('woman')
        coded_salient_items.append('1')
    elif word in umbrella:
        #salient_items_found.append('umbrella')
        coded_salient_items.append('1')
    elif word in flowers:
        #salient_items_found.append('flowers')
        coded_salient_items.append('1')
    elif word in clogs:
        #salient_items_found.append('clogs')
        coded_salient_items.append('1')
    elif word in leash:
        #salient_items_found.append('leash')
        coded_salient_items.append('1')
    elif word in rain:
        #salient_items_found.append('rain')
        coded_salient_items.append('rain')
    elif word in puddle:
        #salient_items_found.append('puddle')
        coded_salient_items.append('1')
    elif word in clock:
        #salient_items_found.append('clock')
        coded_salient_items.append('2')
    elif word in coat:
        #salient_items_found.append('coat')
        coded_salient_items.append('1')
    elif word in grass:
        #salient_items_found.append('grass')
        coded_salient_items.append('grass')
    elif word in path:
        #salient_items_found.append('path')
        coded_salient_items.append('1')
    elif word in walk:
        #salient_items_found.append('walk')
        coded_salient_items.append('1')
    elif word in tower:
            #salient_items_found.append(’tower’)
            coded_salient_items.append(’2’)




    ##I have inserted a blank at the 1st
    ##and last position, such that every
    ##item in list is counted twice.
    blank = "...."
    coded_salient_items.append(blank)
    coded_salient_items.insert(0,blank)
    position = 0
    bigram_windows = []

    for item in coded_salient_items:
        if position+1 < len(coded_salient_items):
            bigram = coded_salient_items[position] + "-"
                        + coded_salient_items[position+1]
            bigram_windows.append(bigram)
            position = position + 1

    #print dutch_match_words
    return bigram_windows

def count_bigram(file_list):
    "Counts different types of transitions
         in the bigram pairs."
    bigram_sum = 0.0
    total_file_bigrams = len(file_list)
    norm_total = file_list.count(’1-1’) +
        file_list.count(’2-2’) +
        file_list.count(’....-1’) + file_list.count(’1-....’) +
        file_list.count(’....-2’) + file_list.count(’2-....’)
    transitional_total = file_list.count(’1-2’) +
        file_list.count(’2-1’)
            + 0.0
    transitional_non_region = file_list.count(’rain-1’)
            + file_list.count(’1-rain’) +
        file_list.count(’rain-2’) +
        file_list.count(’2-rain’) +
        file_list.count(’grass-1’) +
        file_list.count(’1-grass’) +
        file_list.count(’grass-2’) +
        file_list.count(’2-grass’)
    same_non_region = file_list.count(’rain-rain’)
            + file_list.count(’grass-grass’) +
        file_list.count(’rain-....’) +
        file_list.count(’....-rain’) +
        file_list.count(’grass-....’) +
        file_list.count(’....-grass’)
    transitional_percent = transitional_total/len(file_list)

    #print file_list
    print str(transitional_percent) + ’,’+ str(transitional_total) +
         ’,’ + str(norm_total) + ’,’ + str(transitional_non_region)
         + ’,’ +
          str(same_non_region) + ’,’ + str(total_file_bigrams)
    #print transitional_percent
```