Data mining brain regions with neuroimaging databases

Helgi Már Sigurðsson

Kongens Lyngby 2012
IMM-M.Sc-2012-88

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-M.Sc-2012-88

Abstract

This thesis presents methods of automatic meta-analysis of neuroimaging articles that report brain region coordinates. Two lists were gathered, a list of brain regions and a list of cognitive terms. 1.646 articles were downloaded and the coordinates in them were linked to the brain regions in the list that had been gathered, which made the articles connected to the brain regions as well. For each brain region the abstracts of all the articles that had been linked to it were data mined with the list of cognitive terms. Non-negative matrix factorization was then used to discover topics in each of those brain regions. Each topic had a certain set of articles linked to it, and thus also a certain set of coordinates. The topics in each brain region were then compared with a statistical test to see if the distributions of their respective coordinates were similar or not. If the statistical test showed that the distributions were dissimilar that could mean that two topics within the same brain region were functionally segregated. The results were then compared to results from previous studies of chosen brain regions.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark in fulfilment of the requirements for acquiring an M.Sc. in Computer Science and engineering.

The thesis deals with automatic meta-analysis of neuroimaging articles to find topics within brain regions. The topics are then compared with a statistical test to see if different topics are functionally segregated within a single brain region.

The thesis supervisors are Finn Årup Nielsen and Lars Kai Hansen.

Lyngby, 29-February-2012

Helgi Már Sigurðsson

Acknowledgements

I would like to thank my supervisor Finn Årup Nielsen for his guidance, sound advice, our many meetings and for letting me do the project part-time in Iceland. Special thanks go out to David Van Essen for sharing the SumsDB database with me. I would also like to thank my parents, Guðrún and Diddi, for their endless support and letting me stay with them while in Iceland. I wish to thank my good friend Nonni Bergmann for all our discussions on Python and mathematics. Lastly I would to thank my oldest friend, Jónsi, for all his helpful comments.

Contents

Chapter 1

Introduction

The number of functional neuroimaging articles published per year is very large and ever increasing. Articles describing results from functional magnetic resonance imaging (fMRI) procedures have for example grown since 1992 from just a few articles to being almost 2.500 in the year 2005[17]. These articles report coordinate-based results where the coordinates denote where the brain is activated when performing a particular task. Each article might report anywhere from a single coordinate to a few hundred coordinates. With a large set of neuroimaging articles and coordinates comes the possibility of data mining and finding patterns and knowledge that a single article would not reveal. There are a number of projects that are doing exactly that.

BrainMap[1] is a project that started in 1988 and has a database with over 2.000 functional neuroimaging articles that have more than 80.000 coordinates[8]. They have developed software and tools to enable meta-analysis and data mining of the articles as well as distributing software and concepts for quantitative integration of neuroimaging data.

Surface Management System Database, or SumsDB[2], is a repository of brain mapping data from many laboratories, with over 2.000 neuroimaging articles and just over 150.000 coordinates[3]. SumsDB allows for flexible search options

[1] http://www.brainmap.org
[2] http://sumsdb.wustl.edu:8081/sums

of coordinates and they have created software that enables visualizations of the coordinates on a human brain atlas[23].

Neurosynth[3] is a framework that uses text mining and meta-analysis to find mappings between brain activity and cognitive states[29]. They have more than 4.000 articles and nearly 150.000 coordinates in their database. Their website allows for visualization of coordinates on a brain template. Coordinates can be viewed for a certain term or a topic, where the topics have been found using Latent Dirichlet Allocation (LDA)[4].

brainSCANr[4] is a website that has collected brain region names, cognitive and behavioural functions and disease names to find out how often any two phrases appear in the same articles[25]. They have analysed the text of more than 3.5 million scientific abstracts and their assumption is that the more often two terms appear in the same article the more likely they are to be associated. Graphs can be viewed were the associations between terms can be seen.

In 2005 Nielsen et al. described a method for automatic meta-analysis of neuroimaging articles[12]. They downloaded abstracts from PubMed[5], concerning the posterior cingulate cortex (PCC), and turned them into a bag-of-words. The data was then analysed with non-negative matrix factorization[10] to discover latent classes ('topics'). The distributions of coordinates from two different topics were then analysed with statistical tests. Their findings suggested a functional segregation between memory and pain in the PCC.

This thesis builds on the methods described by Nielsen et al.[12] but with some variations. A list of brain regions and cognitive terms is gathered, and neuroimaging articles are downloaded from SumsDB[3] and the Brede database[13]. The coordinates in these articles state which brain region, or regions, they occur in and thus each article is linked to one or more of the brain regions in the list. For each brain region all the articles that had been linked to it are retrieved and instead of turning them into a bag-of-words like Nielsen et al. did a search is made in their abstracts for the cognitive terms that had been gathered. A document-term matrix is created where each index in the matrix shows how many times a certain term is found in a given abstract. The document-term matrix is then factorized into two matrices with non-negative matrix factorization, which identifies topics for each brain region. The topics in each brain region are then compared with a statistical test to try to find out if different topics are functionally segregated within a brain region. To see if this method actually works the results will be compared to previous studies done in the posterior cingulate cortex, anterior cingulate cortex and the superior temporal sulcus.

[3] http://neurosynth.org/
[4] http://www.brainscanr.com/
[5] http://www.ncbi.nlm.nih.gov/pubmed/

Chapter 2

Design

This project required a database for all the data that was gathered and created, a program for the data mining and a website where the data and results could be viewed. This chapter briefly explains the architecture of the programs made for the data mining, the website where the results are displayed, as well as the database that was created.

2.1 Data processing

The program that was made for the data processing and data mining was written in the Python programming language. Figure 2.1 displays a high-level overview of the Python classes that were created. The classes can be split into two levels, the database level and the business level. The classes dbConnection and dbCommands are on the database level and all the other classes are on the business level. Below each class is described in short detail.

- **dbConnection** stores the connection to the database
- **dbCommands** connects to **dbConnection** and has all the database queries and commands.

- **BrainRegions** parses the list of brain regions that was gathered, maintaining the brain region hierarchy, and sends it to the database level.
- **Sumsdb** and **Brede** handle the articles gathered and take care of sending all the necessary data from them to the database level.
- **Transformations** is used for transforming all the coordinates into the same stereotaxic space.
- **DataMining** is used for finding cognitive terms in the article abstracts.
- **NMF** stands for Non-negative matrix factorization and it takes care of finding topics within each brain region.
- **DistributionTest** compares the distributions of coordinates from two separate topics.

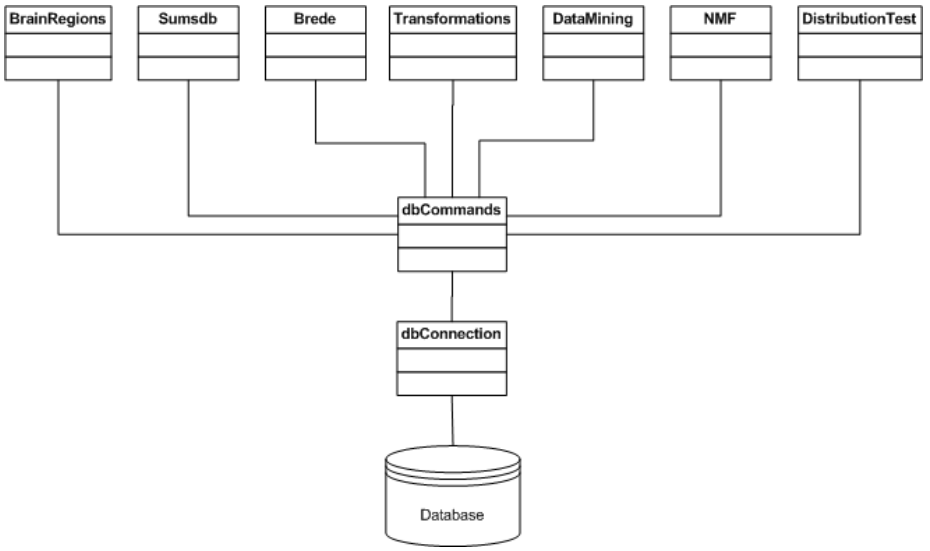


Figure 2.1: Overview of the structure for data processing

2.2 Website

In order to visualize the data that was gathered and created in this project a website was created using the PHP web programming language. The website was divided into three main pages: Articles, Brain Regions and Cognitive Terms.

- On the Articles page all the articles in the database are shown in a list, and they can be filtered by either title, authors or PubMed Id. Each article can then be viewed to show more details about it.
- The Brain Regions page shows all the brain regions in the database. For each brain region the information that is shown is its parent and child regions, the topics found for it, the results of the topic comparisons, the cognitive terms found for it and how many times each one was found, and all the articles that were linked to it.
- The Cognitive Terms page shows all the cognitive terms in the database and for each term it can be seen which brain regions it was associated with and which article abstracts it was found in.

The website can be viewed at http://brainiac.adolf.is

2.3 Data storage

This project required large amounts of data. Some of the data was gathered from outside sources, i.e. articles, coordinates, brain regions and cognitive terms. Then there was the data that came as a result of the topic mining and spatial mining, namely topics and topic comparisons. To store all the data a relational MySQL database was created. An entity-relationship (ER) diagram for the database can be seen in Figure 2.2. Following is a brief description of the tables in the database.

- **articles, articleMesh, mesh, articleAuthors, authors, articleCoordinates** and **coordinates** store all the information that was gathered from individual articles.
- **cognitiveTerms** holds the list of words that were used to identify topics in the brain regions. **articleCognitiveTerms** stores which cognitive terms appear in which article abstracts, and how many times.
- **brainRegions, nameVariations** and **abbreviations** contain the information about all the brain regions, including variations and abbreviatons of the brain region names.
- **brainRegionHierarchy** stores the parent/child hierarchy for the brain regions.
- **topics** keeps hold of the topics that were found for each brain region.
- **topicArticles** and **topicCognitiveTerms** tell which articles and cognitive terms belong to a topic.
- **spatialMining** has the results from the topic comparisons.

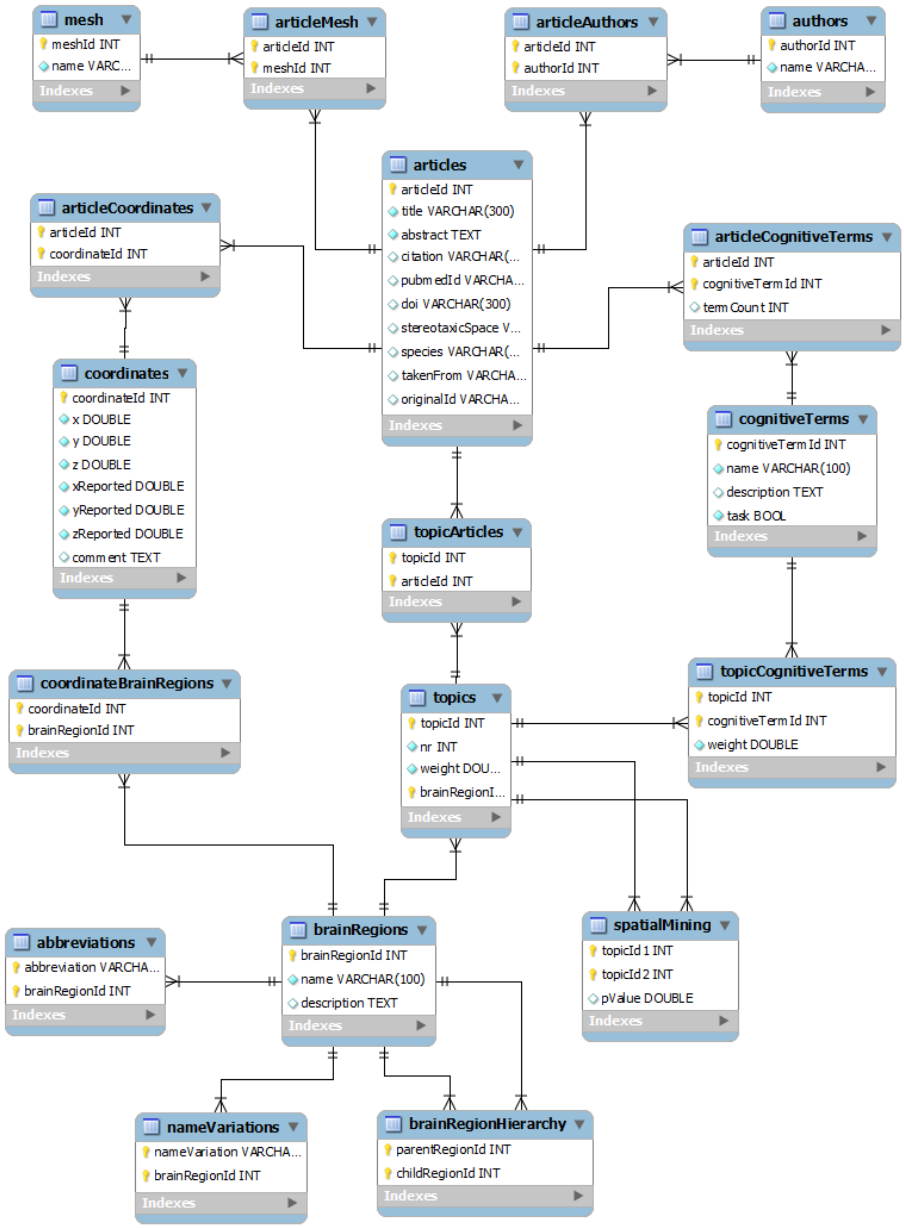


Figure 2.2: Database schema

Chapter 3

Data collection and preprocessing

Before any data mining could be done a lot of data had to be collected. First of all a hierarchical list of brain regions was needed, then a good list of cognitive terms, and lastly as many neuroimaging articles that report brain coordinates as possible. This chapter tells were these data were taken from and describes how the coordinates were transformed so they would all be in the same stereotaxic space.

3.1 Brain regions

The human brain is divided into hundreds of regions, many of which are subregions of other regions. A great deal of brain regions are known under more than one name, the *Posterior cingulate gyrus* is for example sometimes simply referred to as the *Posterior cingulate*, and other times it is abbreviated to *PCgG*. It was therefore vital to the project to have a good hierarchical list of brain regions, with variations and abbreviations of the names.

The brain region list used in this project was taken from the Brede database[13]. The Brede database has a total of 763 regions with 491 name variations and 390 abbreviations. Then another 28 regions, with 13 name variations and 21 abbreviations, were added from the Brede wiki[14]. In total there were 791 brain regions added to the database, with 504 name variations and 411 abbreviations. This is not a complete list of brain regions but for the purpose of this project this was considered a large enough list.

3.2 Cognitive terms

It was decided to use a predetermined list of cognitive terms to data mine the article abstracts with. There are a number of websites that have good lists of cognitive terms. The Brede database[13] for example has one, but since the brain region list had already been taken from there it was decided to use different sources for the list of cognitive terms in order to maintain diversity and not rely too heavily on one source for information.

The list of cognitive terms was taken from two websites, Cognitive Atlas[1] and CogPO[2].

Cognitive Atlas is a project that aims to build an ontology of cognitive terms and mapping them to brain systems[18]. Their database of terms is constantly growing and the project is still in development, but when the terms were downloaded for this project their list counted 887 cognitive terms.

CogPo is a website with a similar agenda to Cognitive Atlas, i.e. to develop an ontology of cognitive terms[1]. Their list of terms is a lot smaller than the one from Cognitive Atlas, when it was downloaded for this project it had 81 cognitive terms. Most of these were actually already in Cognitive Atlas, but still 9 additional cognitive terms were added from CogPo.

In total there were 896 cognitive terms added to the database from the two previous mentioned websites.

[1] http://www.cognitiveatlas.org
[2] http://www.cogpo.org

3.3 Articles

The most important part of the data collection was to find and collect as many neuroimaging articles and brain region coordinates as possible. There are a few websites that have gathered articles and coordinates and in this project they were taken from two of those, namely SumsDB and Brede. Below each of those two websites are discussed in a brief manner.

3.3.1 SumsDB

Surface Management System Database, or SumsDB, is a repository of brain mapping data, including neuroimaging articles and coordinates, from many laboratories[3]. At the time of this writing the SumsDB website states to have 2.344 neuroimaging articles in its database and just over 150.000 coordinates. David van Essen at SumsDB was kind enough to provide the SumsDB database of articles and coordinates in two XML files, one for the articles and one for the coordinates[24]. The database received from David van Essen was quite a bit smaller though than what is reported on the website, a total of 1.619 articles reporting 52.254 coordinates. A few of those articles were however not written about humans and were therefore not reporting coordinates in the human brain, those were excluded leaving 1.593 articles reporting 52.029 coordinates.

Each coordinate in the SumsDB database has a field called *geography* and that field tells which brain region, or regions, the specific coordinate lies in. Even though the brain region list from the Brede database has both variations for names of the brain regions and abbreviations a lot of the regions in the SumsDB file have different names. Some of this is of course due to the fact that not every single brain region is in the Brede database, but in many cases the brain regions in SumsDB were simply spelled differently. Because of this, modifications of the brain region names in the SumsDB file were made so that the coordinates would be matched to one of the brain regions taken from the Brede database. Also a list of the most common brain regions found in the SumsDB file but not found in the Brede database was gathered, and this led to the 28 extra brain regions being added from the Brede wiki (as mentioned in Section3.1). Out of the 52.029 coordinates taken from SumsDB 48.853 of them got connected to one or more of the brain regions in the database.

3.3.2 Brede

The Brede database is freely available for download as a single file[13]. It is not nearly as large as SumsDB and some of the articles in it are the same as in SumsDB. When the data was taken from the Brede database for this project it had 186 neuroimaging articles with 3.912 coordinates. Out of those 186 articles 53 of them, reporting 1.102 coordinates, had not been found in SumsDB so they were added to this project's database.

Of the 1.102 coordinates added from the Brede database 861 one of them got connected to one or more brain regions in the database.

3.4 Transforming coordinates

Brain region coordinates can be reported in a number of different stereotaxic spaces, the most common ones being the Talairach coordinate space[20] and the MNI coordinate space[2]. It was important for the statistical tests to have all the coordinates in the same stereotaxic space, if possible. From the coordinates that had been gathered most of them were already in the Talariach space. It was therefore decided to try to have all the coordinates in the Talairach space by using known transformations between stereotaxic spaces.

The coordinates in the Brede database are kept in both their original stereotaxic space and the Talairach coordinate space. If the original space is not the Talairach space they are transformed into it and therefore the coordinates from the Brede database required no transformations.

The coordinates from SumsDB are however only kept in their original stereotaxic space and therefore transformations were needed. The following stereotaxic spaces were found in the articles from SumsDB:

- 711-2B
- 711-2C
- 711-2Y
- AFNI
- FLIRT
- MRITOTAL

- Other
- SPM2
- SPM5
- SPM95
- SPM96
- SPM99
- T88

Some of these are in fact not stereotaxic spaces but rather the software used to register the coordinates in a certain stereotaxic space, but that software can indicate which stereotaxic space the coordinates are in.

The transformations used in this project are the same as used by BrainMap[3], which is the Lancaster transform (i.e. icbm2tal), where coordinates are transformed between the MNI and Talairach spaces[9]. In BrainMap they have three different ways of converting between MNI and Talairach spaces, depending on the software used for deriving the MNI coordinates, namely *icbm_spm2tal*, *icbm_fsl2tal* and *icbm_other2tal*. The Matlab code for these transformations is available on the BrainMap website and for this project it was taken and converted to Python code, since Python was the programming language being used.

No known transformation were found for the 711-2B, 711-2C and 711-2Y spaces but they are apparently similar in size to Talairach space[16, 5] so the coordinates in those spaces were left unchanged. There was no way of knowing weather AFNI referred to MNI or Talairach coordinates, it could be both, so they were left unchanged as well. Most the other ones are either in the Talairach space or fell within the usage of the transformations from BrainMap, i.e. *icbm_spm2tal* for SPM2, SPM5, SPM96 and SPM99, *icbm_fsl2tal* for FLIRT and *icbm_other2tal* for Other. MRITOTAL is not mentioned by BrainMap but those coordinates were in MNI space and *icbm_other2tal* was used to transform them. There's no way of knowing whether that is the right thing to do but at least it is better than leaving them in MNI space. SPM95 and T88 refer to coordinates in Talairach space so they were left as they were.

[3] www.brainmap.org/icbm2tal

Chapter 4

Methodology

This chapter describes the methods used in finding the topics for each brain region and how the coordinates for all the different topics within one brain region were compared in order to see if the topics were possibly functionally segregated within that brain region.

4.1 Finding cognitive terms in abstracts

To begin with a search in the abstracts of all the articles was made to see if any of the cognitive terms could be found, and if found how many times they appeared. In order to find as many instances of the relevant terms as possible three different steps were taken when searching in the abstracts. Two things stayed the same though in each of the steps:

1. All the words in the abstracts as well as all the cognitive terms were turned into lower case.

2. The regular expression [\b*word*\b] was used to do the search.

In that regular expression *word* was the cognitive term being searched for and \b indicates a word boundary. Some of the cognitive terms can be found as parts of others words, the word boundary prevents that. An example is the word *action* which without the word boundary can be found e.g. in the word *fraction* and clearly that is not the desired result. The three steps taken when searching for the cognitive terms in the abstracts were as follows.

The first step was the most straightforward one, the regular expression was simply used on the abstracts as they were.

If after the first step the term had not been found the abstract was lemmatized using the WordNet Lemmatizer from Python's NLTK library. Lemmatization is the process of finding the dictionary form (or lemma) of a word[26] and the WordNet Lemmatizer does this by use of WordNet® which is a large database of English words at Princeton University where words are grouped together based on their meanings[21]. An example of when this came in handy in this project is the word *strategies*, a word that was not found before, which changed to its dictionary form *strategy* after the lemmatization. *Strategy* was indeed one of the terms being searched for

If still nothing was found the words in the abstract were stemmed using the Lancaster stemmer from Python's NLTK library. Stemming is related to lemmatization, the difference being that stemmers have no knowledge of the other forms of the word they are examining[27]. The Lancaster stemmer for example doesn't have a dictionary to look up in like the WordNet Lemmatizer, instead it has its own rules for finding the stem of a word. An example of a word the Lancaster stemmer was useful for is *fearful*, which is not one of the terms being searched for but after stemming it becomes *fear*, which is in the list of terms. This would not have been caught by the WordNet Lemmatizer since the word *fearful* is already in its dictionary form.

These three steps of finding cognitive terms in abstracts were used in the order described above because it gave the best results. Performing them in any other order resulted in less terms being found. This method does however not find every single instance of the terms it is looking for and some instances are surely missed. More instances could have been found e.g. by doing the above mentioned steps on top of each other, instead of only doing steps two and three if the step before did not find anything. This would have caught some of the missed terms, but it would also have counted some of the terms more than once, resulting in more terms being found than were actually in the abstract. It was decided that it was better to find fewer terms than to find more terms than were present.

Out of the 896 cognitive terms in the database 422 of them were found a total of 13.764 times, which means that each of the terms found was on average found just under 33 times. Out of the 1.646 articles in the database one or more of the cognitive terms were found in the abstract of 1.581 of them, meaning that on average approximately 4.3 terms were found in each abstract.

After this was done each of the cognitive terms found was linked to one or more articles. Having certain articles linked to a cognitive term gave the possibility of linking that term to a set of coordinates. As an example a visualization of all the coordinates for the term *fear* can be seen in Figure 4.1. Also since the coordinates of the articles had previously been linked to the brain regions in the database it was now possible to see for each cognitive term found which brain regions it was being mentioned in the most. An example of this for the same word, *fear*, can be seen in Figure 4.2 where the two most common brain regions and their hierarchy can be seen. A full list of brain regions for each term can be seen on the website at http://brainiac.adolf.is/?page=cognitiveterms. Even though this was not the intended purpose for finding the cognitive terms in the abstracts it gave for an interesting by-product.

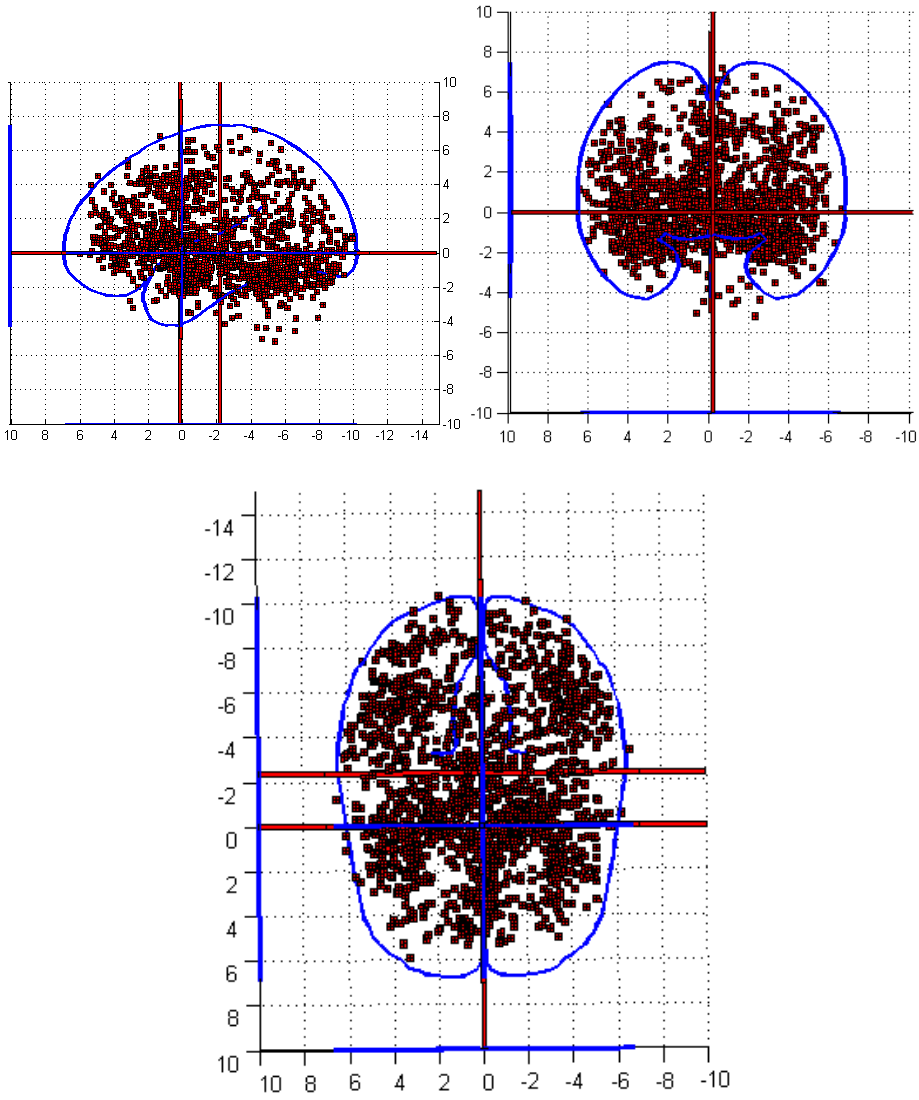
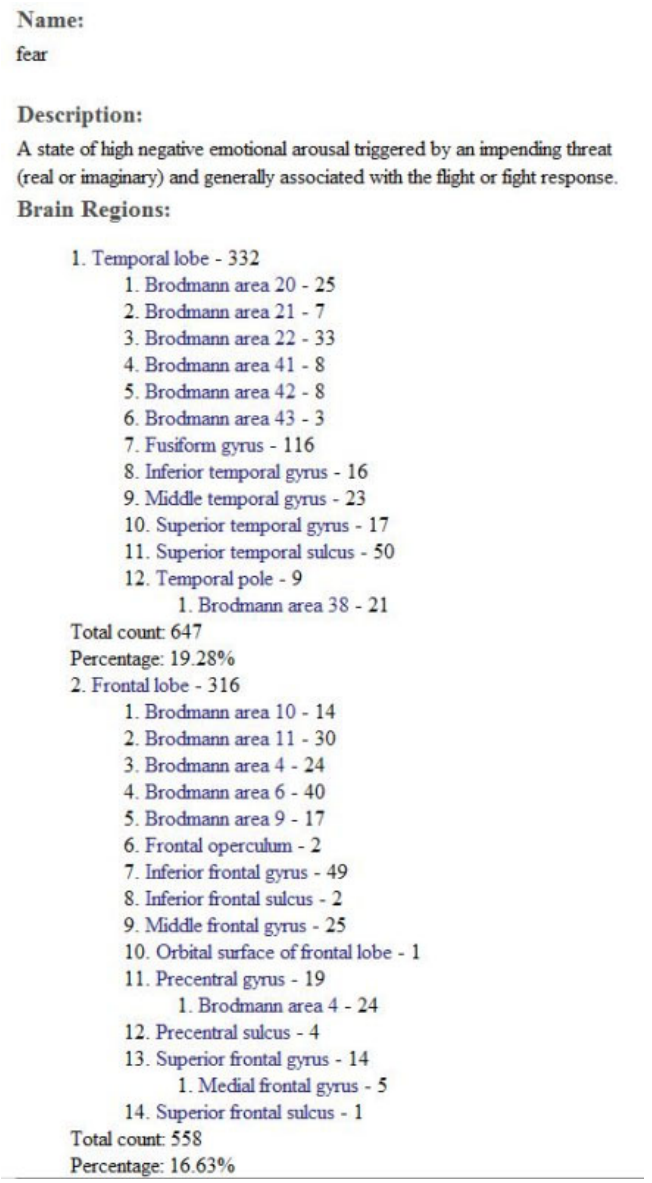


Figure 4.1: All the coordinates found for the term *fear* seen from the left, front and above.



Name:

fear

Description:

A state of high negative emotional arousal triggered by an impending threat (real or imaginary) and generally associated with the flight or fight response.

Brain Regions:

1. Temporal lobe - 332
    1. Brodmann area 20 - 25
    2. Brodmann area 21 - 7
    3. Brodmann area 22 - 33
    4. Brodmann area 41 - 8
    5. Brodmann area 42 - 8
    6. Brodmann area 43 - 3
    7. Fusiform gyrus - 116
    8. Inferior temporal gyrus - 16
    9. Middle temporal gyrus - 23
    10. Superior temporal gyrus - 17
    11. Superior temporal sulcus - 50
    12. Temporal pole - 9
        1. Brodmann area 38 - 21

Total count: 647

Percentage: 19.28%

2. Frontal lobe - 316
    1. Brodmann area 10 - 14
    2. Brodmann area 11 - 30
    3. Brodmann area 4 - 24
    4. Brodmann area 6 - 40
    5. Brodmann area 9 - 17
    6. Frontal operculum - 2
    7. Inferior frontal gyrus - 49
    8. Inferior frontal sulcus - 2
    9. Middle frontal gyrus - 25
    10. Orbital surface of frontal lobe - 1
    11. Precentral gyrus - 19
        1. Brodmann area 4 - 24
    12. Precentral sulcus - 4
    13. Superior frontal gyrus - 14
        1. Medial frontal gyrus - 5
    14. Superior frontal sulcus - 1

Total count: 558

Percentage: 16.63%

Figure 4.2: Screenshot from the website showing the first two brain regions associated with fear, and the subregions in their hierarchy that were also associated with fear.

4.2 Topic mining

This section describes how non-negative matrix factorization was used to discover latent classes in the abstracts for each brain region. The latent classes correspond to what is referred to as topics in this thesis.

4.2.1 Preparing the data

The hierarchy for each brain region in the database was fetched, i.e. its children, children's children etc. When the hierarchy for a certain brain region had been collected all the articles linked to any of the regions in the hierarchy were fetched. For those articles all the cognitive terms, which as described in section 4.1 had previously been found for them, were also fetched. If a term only occurred once in the collection of articles it was removed. Then all the articles and the remaining cognitive terms were converted into a document-term matrix: $M(N \times Q)$ where N was the number of articles and Q was the number of cognitive terms found in those articles. Each element M_{nq} in the matrix had the value of how many times a term Q occurred in article abstract N.

Some terms could be found numerous times in a single abstract, but not found in any (or few) other abstracts. Others could be found in many abstract but very seldom in each one. In order to better represent the values of the cognitive terms to the collection of abstracts a tf*idf weight (term frequency-inverse document frequency) was applied to the matrix. tf*idf is a numerical statistic which reflects how important a word is to a document in a collection or corpus[28].

tf*idf is calculated as:

$$\text{tf*idf}(t, d, D) = \text{tf}(t, d) * \text{idf}(t, D), \tag{4.1}$$

where t is the cognitive term, d is the abstract and D is all the abstracts.

The tf part of the equation, $\text{tf}(t, d)$, is simply the number of times a given term appears in an abstract, which had already been found as explained in section 4.1.

The idf part is then found with the following equation:

$$\text{idf}(t, D) = \frac{|D|}{|d \in D : t \in d|}, \tag{4.2}$$

with $|D|$ as the total number of abstracts for a brain region and $|d \in D : t \in d|$ as the number of abstracts a term was found in.

The denominator is often adjusted to $1 + |d \in D : t \in d|$ to avoid division-by-zero, in case a term would not be in any of the abstracts. In this project that was not necessary since only the terms that had been found in the abstracts were fetched.

4.2.2 Non-negative matrix factorization (NMF)

Non-negative matrix factorization, or NMF, was used to discover topics within the brain regions. There are several algorithms that implement NMF, but in this project the algorithm used is the one described by Lee et al. in 2001[11].

NMF factorizes a non-negative matrix $X(N \times Q)$ into two non-negative matrices $W(N \times K)$ and $H(K \times Q)$ such that:

$$X = WH + E, \tag{4.3}$$

where E is the cost function which is defined as the Euclidian distance between X and WH,

$$\|X - WH\|^2. \tag{4.4}$$

The document-term matrix that was described in section 4.2.1 was the matrix that was sent into the NMF algorithm, which makes it X in Equation 4.3. The K in the W and H matrices is the number of topics found for a brain region and that number was decided depending on the number of articles and cognitive terms found for a brain region.

When the W and H matrices were created they were initialized with random values. This could result in slightly different results each time the NMF algorithm was run with the same data, but not significant enough to make a great difference. Initializing the matrices with random variables is the only viable option since doing it any other way could lead to the calculations being biased towards a specific result. The algorithm was run iteratively with an upper boundary of 5.000 iterations. With each iteration new values for W and H were calculated with the hope of reaching a local minimum. It would of course be optimal to find a global minimum but according to Lee et al. that is unrealistic[11]. Since the algorithm was being run over hundreds of brain regions it would have taken a very long time to find local minima for each of them. So for the sake of optimization the iterations were stopped when the change in the cost function value between iterations went below a certain threshold. Even though that means that a local minimum would not have been reached, it should at least have taken us fairly close to it. When the threshold had been reached the W and H matrices were returned. The threshold was determined by the following Python code:

```
max(X.shape) * 1000 * numpy.finfo(float).eps,
```

where numpy.finfo(float).eps is the smallest representable number in Python's Numpy library such that $1.0 + \text{eps} \neq 1.0$[15], and max(X.shape) returns the largest dimension of the X matrix (either the number of articles or the number of terms).

The W and H matrices were recalculated in each iteration with the following equations:

$$H = H\frac{(W^T X)}{(W^T W H)}, \tag{4.5}$$

$$W = W\frac{(X H^T)}{(W H H^T)}. \tag{4.6}$$

When the NMF algorithm had finished running the two matrices, W and H, were returned, where W was $articles \times topics$ and H was $topics \times terms$. Each element in the matrices had a value, or weight, depending on how important certain articles or cognitive terms were to a topic. It was desirable that the articles and cognitive terms only belong to a single topic. In order to get that result a winner-take-all function was applied to the matrices where in W the highest value in each row was kept whereas all others were turned to 0 and in

H the highest value in each column was kept and all the others were turned to 0. Each topic was then given a weight that corresponded to the total weight of the articles belonging to it.

By doing this it was now possible to find both all the articles and all the cognitive terms belonging to a certain topic, and since the articles had coordinates associated with them each topic now had its own collection of coordinates.

4.3 Spatial mining

Having different topics within a brain region and each with its own set of coordinates made it possible to statistically compare their distributions to see if different topics within a brain region were functionally segregated. For each brain region all the topics in it were compared to one another with Hotelling's two-sample T-squared statistic[22].

4.3.1 Hotelling's T-squared distribution

As in the topic mining each brain region and its hierarchy were fetched and for each brain region the topics that had been found for it were also retrieved. When the topics were found with the NMF each of them was assigned certain articles and each of these articles had coordinates in them. Most of the articles report coordinates in more than one brain region, therefore only the coordinates that were linked to any of the brain regions in the hierarchy that was being examined were taken.

The topics were iterated over, where each topic was compared to all the other topics in that particular brain region. In each iteration two matrices were created $Z_1(M_1 \times 3)$ and $Z_2(M_2 \times 3)$, where M_1 and M_2 were the number of coordinates for each topic. The two-sample Hotelling's T^2 test was applied to the Z_1 and Z_2 matrices to test if the distributions of the coordinates in them were the same.

First the unbiased covariance between the two sets of coordinates was found as

$$W = \frac{(Z_1 - \bar{z_1})(Z_1 - \bar{z_1})^T + (Z_2 - \bar{z_2})(Z_2 - \bar{z_2})^T}{M - 2}, \tag{4.7}$$

where $M = M_1 + M_2$ and $\bar{z_1}$ and $\bar{z_2}$ are the means for the two sets of coordinates. The covariance was then used in finding the Mahalanobis distance with the following equation:

$$D^2 = (\bar{z_1} - \bar{z_2})^T W^{-1} (\bar{z_1} - \bar{z_2}). \tag{4.8}$$

This was then transformed into F-statistic in the following way:

$$F = \frac{M_1 M_2 (M - P - 1)}{M(M - 2)P} D^2, \tag{4.9}$$

where P was the dimension of the space, which in this case was P=3 since 3-dimensional coordinates were being compared.

With the F value, the numerator degrees of freedom as P, and the denominator degrees of freedom as M-P-1, a p-value was then found. The purpose of obtaining the p-value was not to reject a certain null hypothesis. But the lower the p-value between two topics the more likely it was that they were happening in different places within a brain region.

Chapter 5

Results

After all the topic mining og spatial mining had been done 2.034 topics had been identified, in 376 of the brain regions, with 7.232 topic comparisons. The topic comparisons with low p-values were the ones of interest since a low p-value could indicate that two topics were functionally segregated within a single brain region. Out of the 7.027 comparisons around 1.800 of them had p-values from 0.001 to 0, which is a very significant value. Some of the slightly higher values could even be considered significant. A low p-value is not a guarantee though that two topics are segregated. In order to see if two topic really were functionally segregated their coordinates needed to be visualized. Doing that for close to a 1.800 comparisons (or more) was by far too large of a task to be fitted within the time frame of this project. Instead of trying to describe the results from every single comparison, selected brain regions and topic comparisons are discussed.

Below the results from the topic mining and topic comparisons in the posterior cingulate cortex (PCC), anterior cingulate cortex (ACC) and the superior temporal sulcus (STS) are described and compared to results from previous studies in those brain regions.

5.1 Posterior cingulate cortex (PCC)

In 2005 Nielsen et al. described similar methods as in this project to data mine the posterior cingluate cortex[12]. They found functional segregation between memory and pain where memory brain activations were mostly in the caudal part and pain brain activations mostly in the rostral part of the PCC. They downloaded 271 abstracts from PubMed[1] and turned them into a bag-of-words, instead of having predetermined words to look for as is done in this project. The most common words they found were 'memory', 'alzheimer', 'visual', 'metabolic', 'retrieval' and 'pain', were 'memory' occurred more than twice as often as the second most common word.

In this project the hierarchy for the PCC counts 10 brain regions and 492 of the articles with 1.224 coordinates were linked to one or more of those regions. The most common words found for the PCC were 'memory', 'learning', 'retrieval', 'recognition', 'reward', 'encoding' and 'pain', with 'memory' being found more than twice as many times as any other word. This is quite similar to the results of Nielsen et al.[12]

The topic mining for the PCC resulted in 11 topics being found. Following is a list of those topics, showing the weight for each topic and the terms with the highest weight in each of them:

- **Topic 1 (2.681):** language (0.4287), auditory (0.3107), action (0.2355), perception (0.1615), decision (0.1493), comprehension (0.1292), reading (0.1007), decision making (0.098), expertise (0.0635), imagery (0.0592), language processing (0.0561)
- **Topic 2 (2.25):** memory (0.5928), working memory (0.5187), maintenance (0.2188), manipulation (0.1247), visual working memory (0.0972), interference (0.088)
- **Topic 3 (1.871):** retrieval (0.6719), encoding (0.5559), recall (0.2888), memory retrieval (0.1292), episodic memory (0.1255), autobiographical memory (0.0667)
- **Topic 4 (1.768):** learning (0.8501), prototype (0.1259), categorization (0.112), rule (0.0828), habit learning (0.0754), habit (0.0699), context (0.0654), knowledge (0.0643), category learning (0.0601)
- **Topic 5 (1.37):** recognition (1.03), familiarity (0.1994), monitoring (0.1122), face recognition (0.0511), recognition memory test (0.0418)

[1] http://www.ncbi.nlm.nih.gov/pubmed/

- **Topic 6 (1.589):** reward (0.9296), anticipation (0.1039), choice (0.0922), uncertainty (0.0882), risk (0.0783)
- **Topic 7 (1.595):** pain (1.2331), cognitive load (0.0771), empathy (0.0489)
- **Topic 8 (1.243):** priming (0.4691), explicit memory (0.3907), consciousness (0.08), intention (0.0604)
- **Topic 9 (1.16):** feedback (0.8543), movement (0.0806), auditory feedback (0.0417), hearing (0.0354)
- **Topic 10 (1.116):** search (0.6437), visual search (0.2033), conjunction search (0.129), remembering (0.0655)
- **Topic 11 (0.627):** humor (0.582), stress (0.033)

Results from the topic comparisons showed that the following were the ones with the lowest p-values.

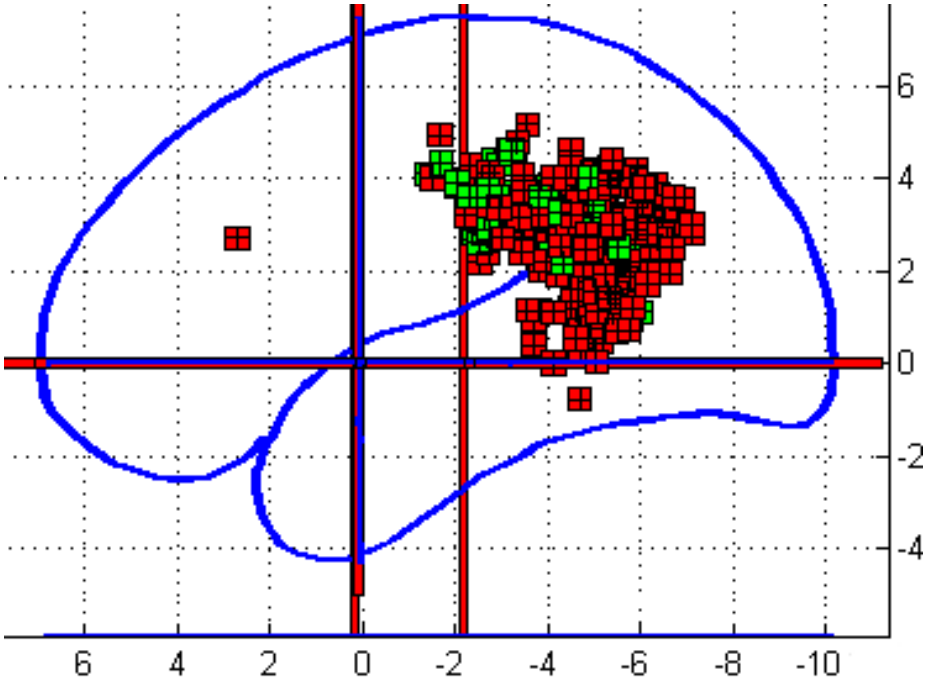
Topic A	Topic B	P-value
retrieval	pain	1.4E-10
recognition	pain	1.5E-8
retrieval	reward	1.6E-5
language	retrieval	2.1E-5
language	pain	2.3E-5
memory	pain	0.0001

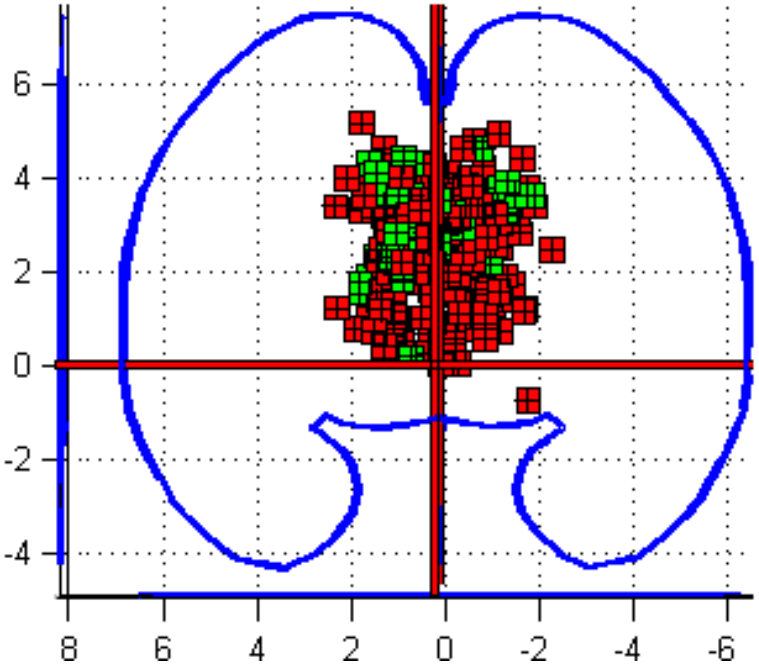
Table 5.1: The lowest p-values from the topic comparisons in the posterior cingulate cortex.

A full list of the terms in each topic and all the topic comparisons can be viewed online at http://brainiac.adolf.is?page=brainregions&id=5.

As can be seen from Table 5.1 the p-value between memory and pain isn't as low as some of the other ones but still low enough to be very interesting.

248 coordinates were found for memory and 78 for pain. Figure 5.1 shows visualizations of the distribution of the coordinates for memory and pain. One coordinate in particular can be seen to be a quite clear outlier and not in the PCC at all. This could bias the results from the topic comparisons. The figure also shows that the memory coordinates tend to be more in the caudal part and the pain coordinates seem to be more in the rostral part of the PCC. This supports the findings of Nielsen et al.[12]


6
4
2
0
-2
-4
6
4
2
0
-2
-4
-6
-8
-10


6
4
2
0
-2
-4
8
6
4
2
0
-2
-4
-6

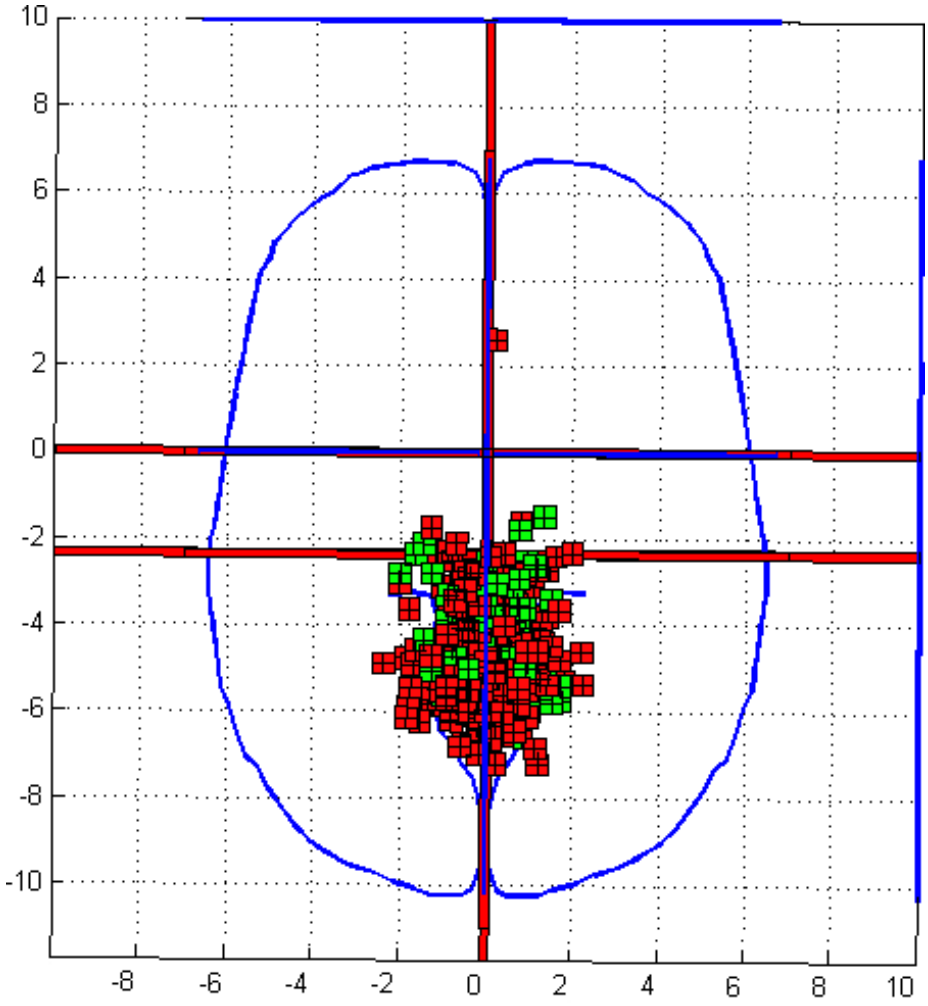


Figure 5.1: The figure above and the ones on the previous page show the distribution of the coordinates for memory, in red, and pain, in green, in the posterior cingulate cortex. Seen from the left, front and above.

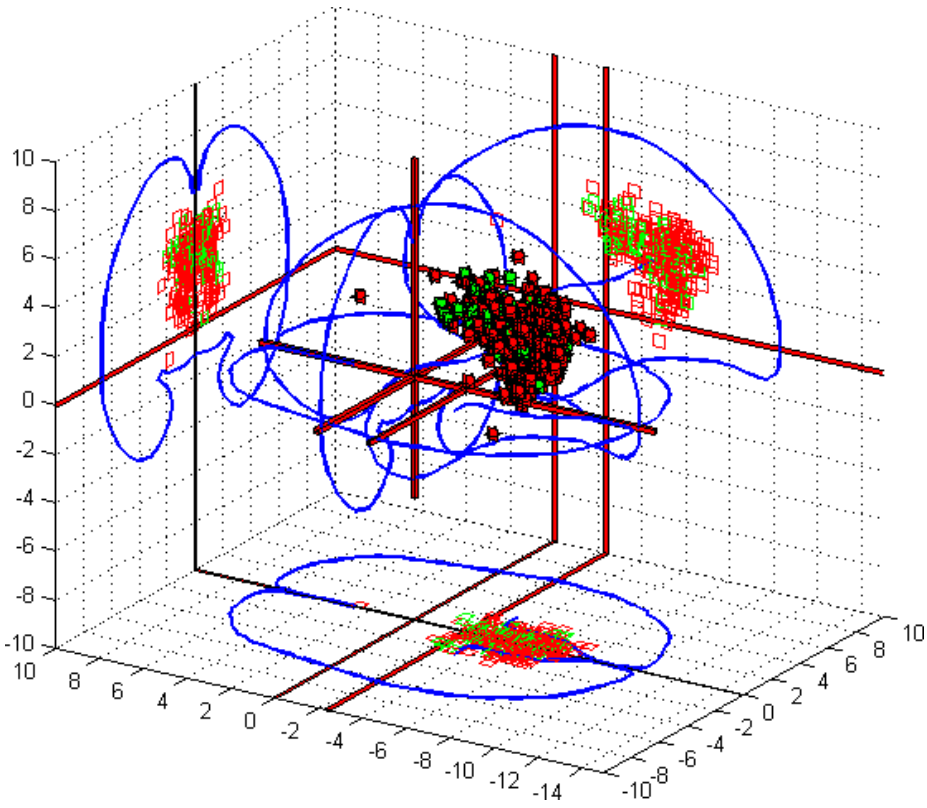


Figure 5.2: A corner cube visualization of memory, in red, and pain, in green, in the posterior cingulate cortex. Memory can be seen to be more in the caudal part and pain in the rostral part of the posterior cingulate cortex.

5.2 Anterior cingulate cortex (ACC)

In the year 2000 Bush et al. discuss the ACC and found that cognitive and emotional tasks happen in separate parts of it[6]. They reported that cognitive tasks are taking place more in the dorsal part while emotional tasks are more in the rostral-ventral part of the ACC.

In the brain region list in this project the ACC has 13 brain regions in its hierarchy. 809 articles with 2.821 coordinates were linked to one or more of those regions. The topic mining for the ACC resulted in 12 topics being found. Following is a list of those topics, showing the weight for each topic and the terms with the highest weight in each of them:

- **Topic 1 (2.745):** memory (1.0262), working memory (0.5529), source memory (0.1735), maintenance (0.1335), rehearsal (0.1197), source memory test (0.0822), distraction (0.0798), attention (0.0708)
- **Topic 2 (2.253):** retrieval (0.8992), encoding (0.6039), episodic memory (0.1896), recall (0.1854), knowledge (0.1373), memory retrieval (0.123), forgetting (0.0932), remembering (0.0876)
- **Topic 3 (2.15):** learning (1.0728), rule (0.1238), surprise (0.1222), prototype (0.1044), integration (0.0898), categorization (0.0699), category learning (0.0677), habit learning (0.0615)
- **Topic 4 (2.118):** decision (0.7507), decision making (0.4158), choice (0.3731), uncertainty (0.165), risk (0.0808), regret (0.0726), action (0.0713)
- **Topic 5 (2.015):** awareness (0.7819), auditory (0.2979), fear (0.2879), arousal (0.2333), consciousness (0.098), valence (0.0721)
- **Topic 6 (2.336):** pain (1.7748), empathy (0.0911), cognitive load (0.0513)
- **Topic 7 (1.729):** reward (1.2745), anticipation (0.2483), error signal (0.0369)
- **Topic 8 (1.519):** recognition (1.0948), familiarity (0.2135), monitoring (0.0962), face recognition (0.064), recognition memory test (0.0472)
- **Topic 9 (1.633):** priming (0.8459), explicit memory (0.2485), implicit memory (0.0608)
- **Topic 10 (1.389):** feedback (0.9451), search (0.157), hearing (0.0664)
- **Topic 11 (1.438):** imagery (0.8236), perception (0.1967), visual imagery (0.1517), movement (0.1462), mental imagery (0.0869), visual perception (0.086)

- **Topic 12 (0.805):** humor (0.6589), language (0.1099), stress (0.0589), discourse (0.0442)

Results from the topic comparisons showed that the following were the ones with the lowest p-values.

Topic A & **Topic B** & **P-value**		
reward	imagery	1.3E-13
reward	memory	5.2E-12
decision	imagery	3.1E-11
reward	pain	4.5E-8
reward	learning	2.4E-6
decision	pain	2.7E-6
decision	memory	6.7E-6

Table 5.2: The lowest p-values from the topic comparisons in the anterior cingulate cortex.

A few other topic comparisons had very low p-values. A full list of the terms in each topic and all the topic comparisons can be viewed online at http://brainiac.adolf.is?page=brainregions&id=8.

It's somewhat difficult to compare these results to what Bush et al.[6] describe since none of the topics that were found can be thought of as typical emotions. This is largely due to the lack of emotional words in the list of words that was mined for. But from Table 5.2 one word stands out quite a bit, Reward, where it can be seen in 4 of the 5 lowest comparisons. Although not classified as an emotion, reward has though been linked to the generation of emotions[19]. Visualisation of the 205 coordinates for reward with the 388 coordinates for memory can be seen in Figure 5.4 and Figure 5.3. The visualizations clearly show that memory lies more in the dorsal part while reward is in the rostral-ventral part of the ACC. This fits within the separation of cognitive and emotional tasks described by Bush et al.[6].

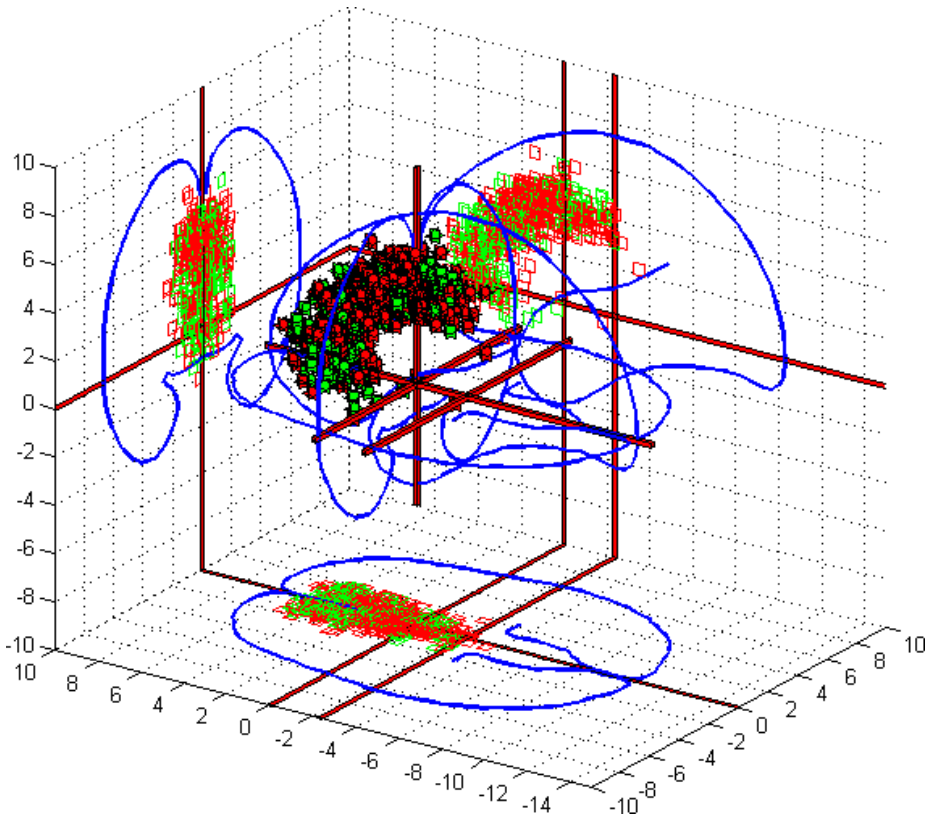
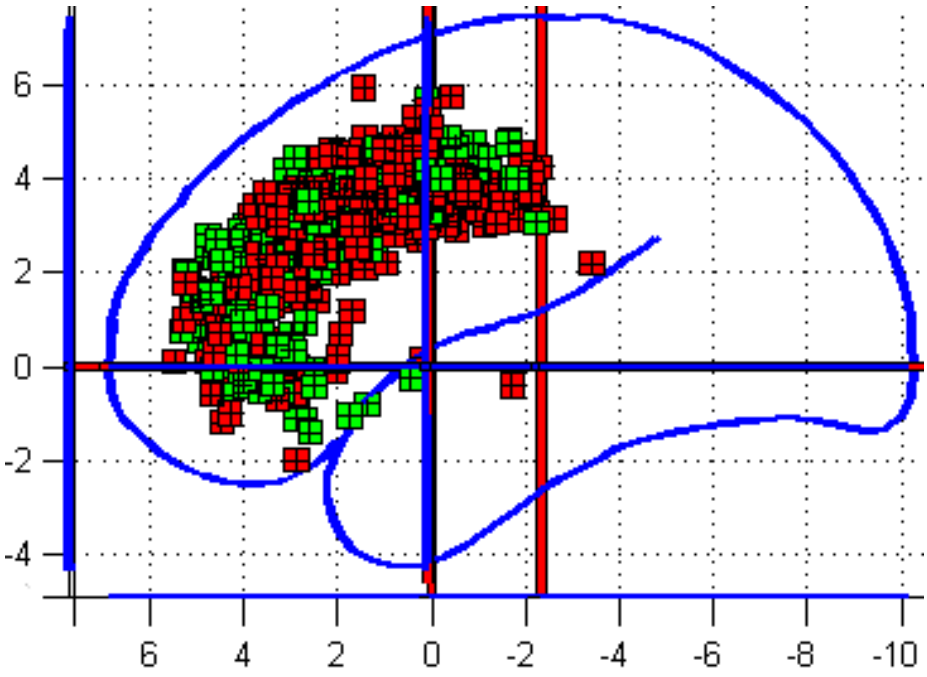
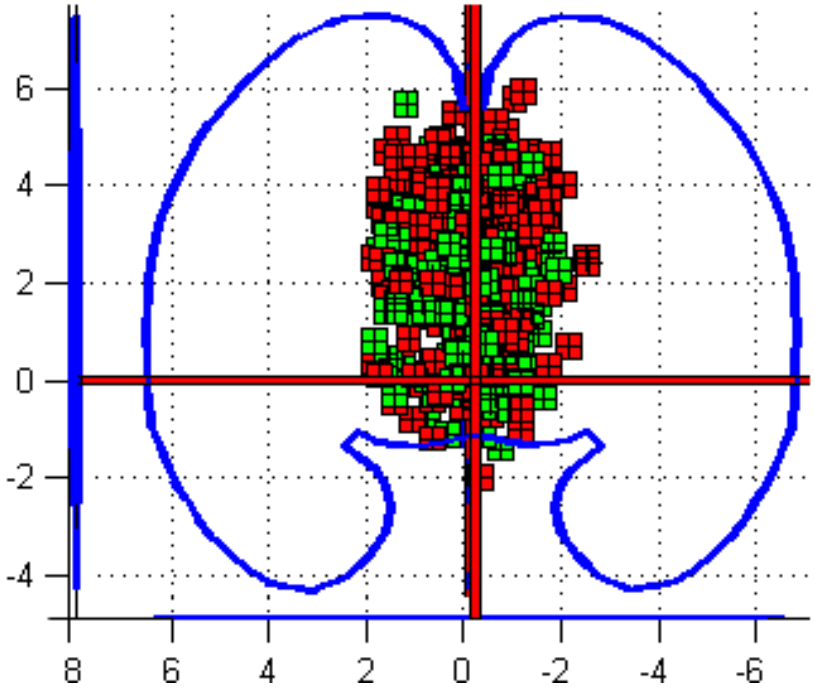


Figure 5.3: A corner cube visualization of memory, in red, and reward, in green, in the anterior cingulate cortex. Memory can be seen to be more in the dorsal part and reward in the rostral-ventral part of the anterior cingulate cortex.


6
4
2
0
-2
-4
6
4
2
0
-2
-4
-6
-8
-10


6
4
2
0
-2
-4
8
6
4
2
0
-2
-4
-6

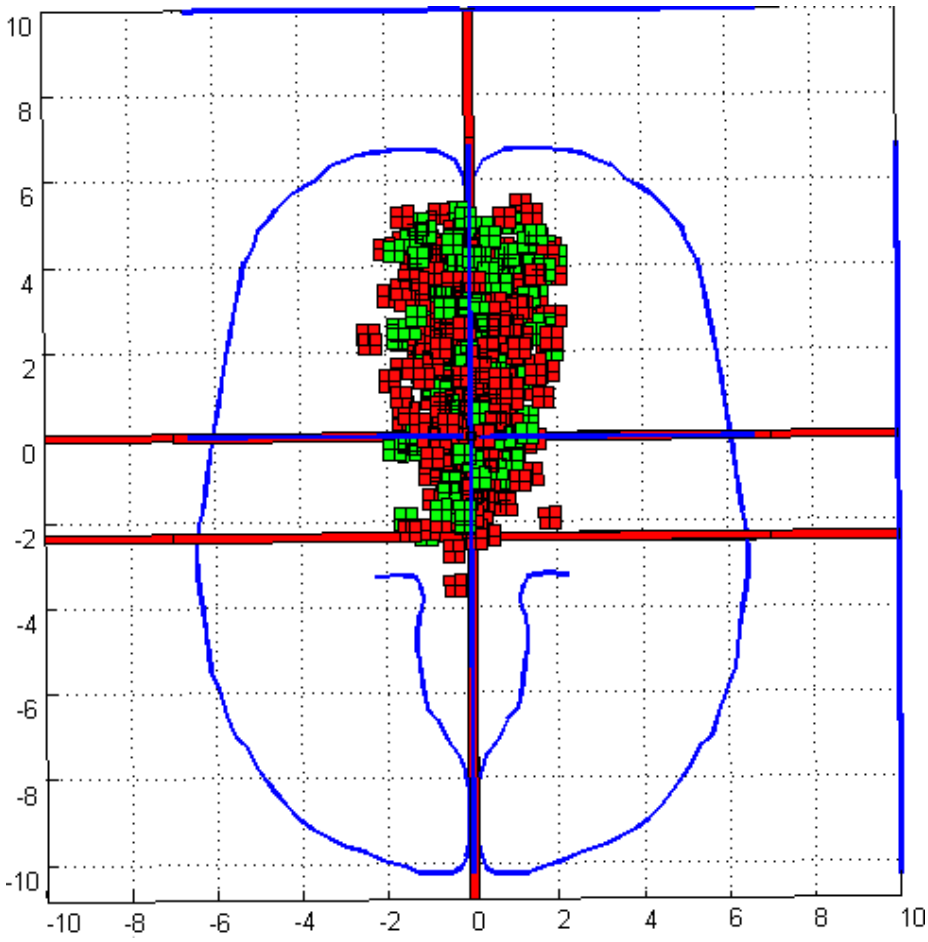


Figure 5.4: The figure above and the ones on the previous page show the distribution of the coordinates for memory, in red, and reward, in green, in the anterior cingulate cortex. Seen from the left, front and above.

5.3 Superior temporal sulcus (STS)

In 2008 Hein et al. describe their results from reviewing coordinates from multiple fMRI studies of the superior temporal sulcus[7]. Their findings revealed distinct clusters of coordinates in the anterior and posterior parts of the STS in both hemispheres. Coordinates for speech processing were mostly found in the anterior part of the STS while coordinates for motion processing, audiovisual integration and face processing where more confound to the posterior part of the STS.

In this project 566 articles with 2.032 coordinates were linked to the STS and 12 topics were identified in it. Following is a list of those topics, showing the weight for each topic and the terms with the highest weight in each of them:

- **Topic 1 (2.619):** language (0.6764), hearing (0.2817), action (0.2537), comprehension (0.1418), integration (0.1206), language processing (0.1121), movement (0.0991), perception (0.0848), awareness (0.0809)
- **Topic 2 (2.254):** memory (0.8734), working memory (0.3617), maintenance (0.1481), manipulation (0.0988), distraction (0.0951), source memory (0.0933), consolidation (0.0851)
- **Topic 3 (1.774):** retrieval (0.7924), knowledge (0.2586), recall (0.1752), forgetting (0.1547), memory retrieval (0.1298), remembering (0.1069), episodic memory (0.1035), autobiographical memory (0.0881)
- **Topic 4 (1.96):** attention (0.8865), auditory (0.2405), audition (0.1356), spatial attention (0.1275), visual attention (0.1217), auditory attention (0.1012), selective attention (0.0849), search (0.0702)
- **Topic 5 (1.462):** recognition (0.9498), familiarity (0.2054), adaptation (0.1342), face recognition (0.0654), face perception (0.0557), object recognition (0.0494)
- **Topic 6 (1.177):** encoding (1.0405), effort (0.0377), multisensory (0.0358)
- **Topic 7 (1.522):** reward (0.7193), risk (0.3269), uncertainty (0.1578), choice (0.0961), decision (0.0663), anticipation (0.0607)
- **Topic 8 (1.293):** priming (0.7035), reading (0.1921), explicit memory (0.1282), expertise (0.0667)
- **Topic 9 (1.281):** feedback (0.8757), monitoring (0.0493)
- **Topic 10 (1.04):** learning (1.1968), association (0.1098)

- **Topic 11 (1.294):** pain (1.0203), cognitive load (0.0887), empathy (0.0624)
- **Topic 12 (0.763):** humor (0.6171), discourse (0.0395), stress (0.0354)

Results from the topic comparisons showed that the following were the ones with the lowest p-values.

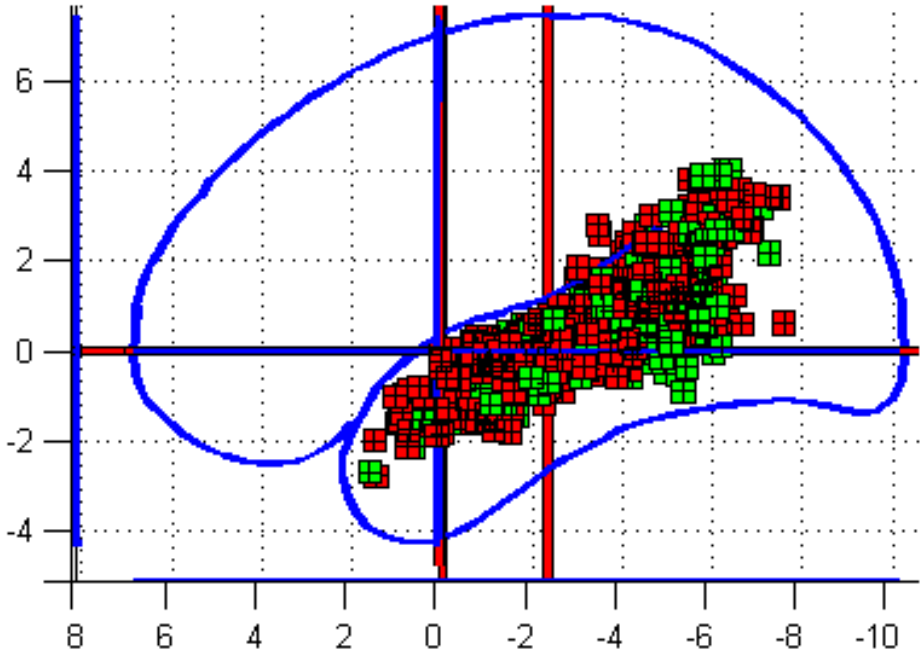
Topic A & **Topic B** & **P-value**		
recognition	attention	4.1E-15
recognition	language	3.5E-14
recognition	memory	8.6E-12
recognition	retrieval	3.1E-11
attention	priming	1.0E-8
attention	language	3.0E-6
recognition	reward	7.3E-6

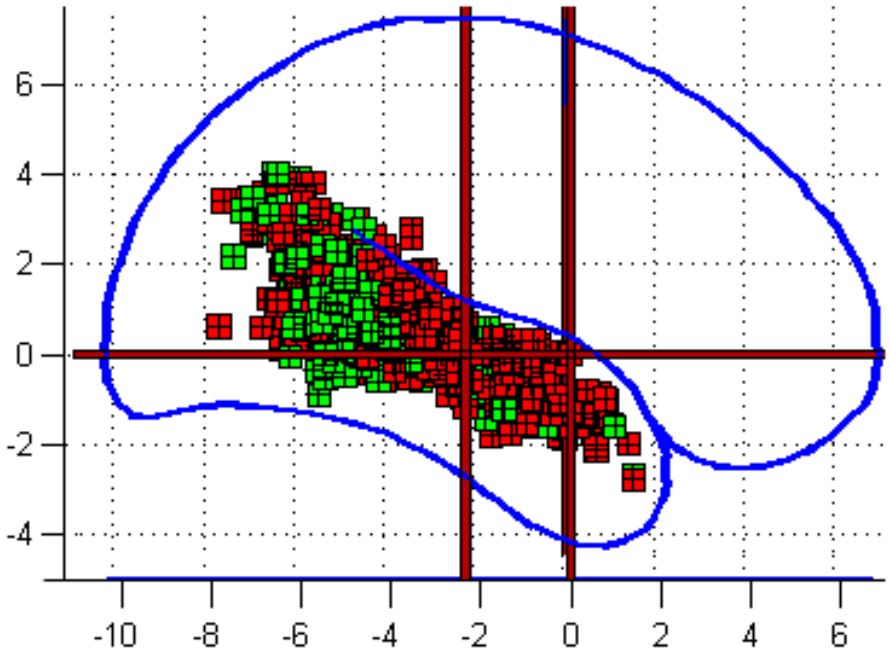
Table 5.3: The lowest p-values from the topic comparisons in the superior temporal sulcus.

Several other topic comparisons had very low p-values. A full list of the words in each topic and all the topic comparisons can be viewed online at http://brainiac.adolf.is?page=brainregions&id=764.

None of the topics found in the STS are an exact match to speech processing, motion processing, audiovisual integration or face processing, which were the topics discussed by Hein et al.[7]. But some of the them are similar and closely related to them. Face processing e.g. was not one of the words in the list of cognitive terms in the database, so that exact topic could not have been found. Topic 5 bares similarities to it though, where the words recognition, face recognition and face perception have a high weight. Topic 1 has similarities to speech processing, with words such as language, hearing, comprehension and language processing having a high weight. Topic 1 (language) and Topic 5 (recognition) were therefore taken as substitutes for speech processing and face processing.

The p-value between these two topics, language and recognition, was 3.5E-14, which is very low. Visualizations of the 622 coordinates for language and 229 coordinates for recognition can be seen in Figure 5.5 and Figure 5.6, where language is represented by red boxes and recognition by green. If recognition is thought of as the same or similar topic as face recognition then the coordinates for it seem to follow the findings of Hein et al.[7] by being clustered more in the posterior part of the STS. The coordinates for language though seem to be a bit more equally distributed across the STS.


6
4
2
0
-2
-4
8
6
4
2
0
-2
-4
-6
-8
-10


6
4
2
0
-2
-4
-10
-8
-6
-4
-2
0
2
4
6

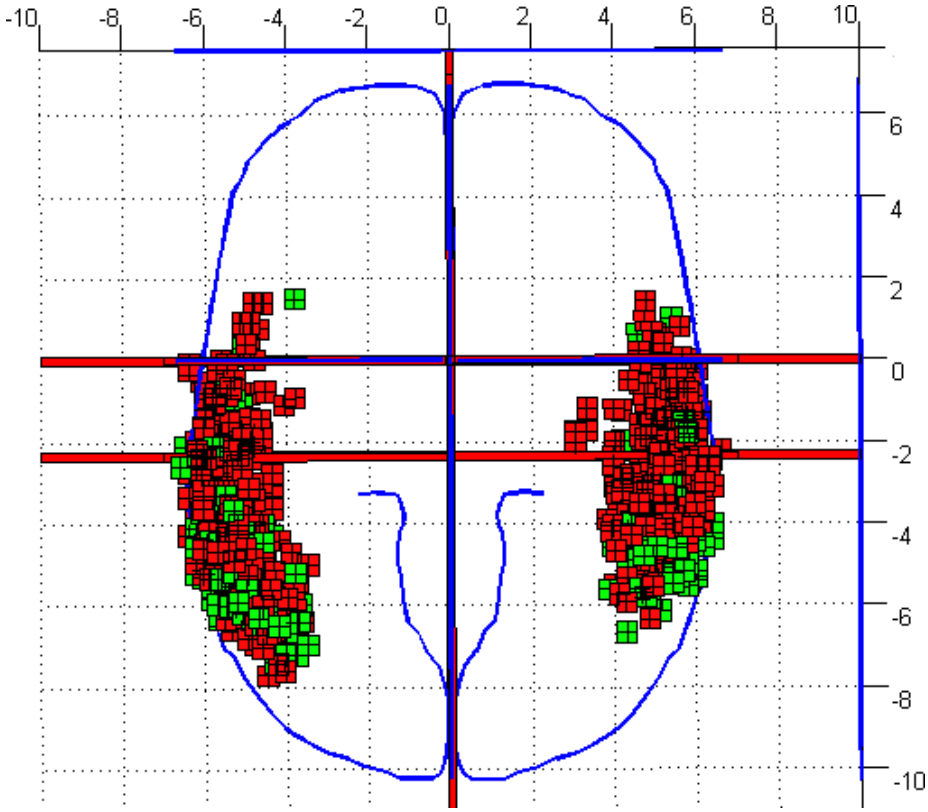


Figure 5.5: The figure above and the ones on the previous page show the distribution of the coordinates for language, in red, and recognition, in green, in the superior temporal sulcus. Seen from the left, right and above.

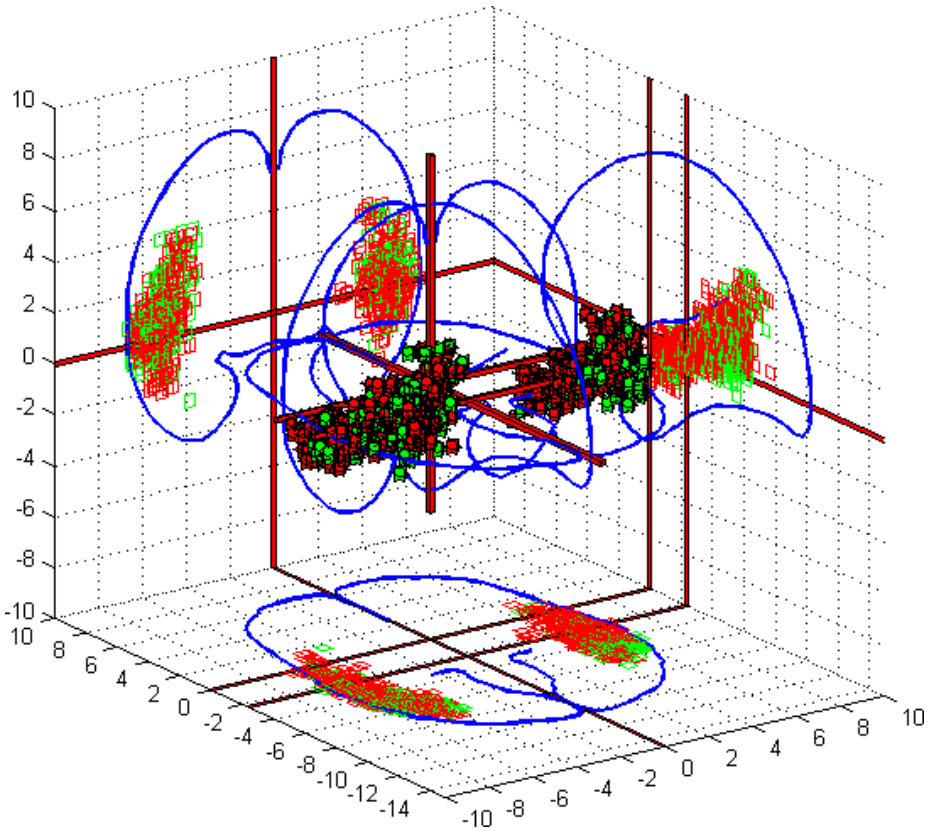


Figure 5.6: A corner cube visualization of language, in red, and recognition, in green, in the superior temporal sulcus. Language seems to be spread all over the STS in both hemispheres while recognition is more confined to the posterior part.

Chapter 6

Conclusion

There are hundreds if not thousands of neuroimaging studies being published every year, each with its own set of coordinates in certain brain regions. The efforts of data mining across multiple studies has already been done in several studies e.g. by Bush et al.[6] and Hein et al[7]. Their methods however involved manual work to summarize their results. With the ever increasing number of neuroimaging studies all manual work will be very time-consuming. Nielsen et al. proposed an automatic method of extracting information about the functions of a certain brain region and applied their methods to the posterior cingulate cortex[12]. This project built on the methods described by Nielsen et al. and used them to data mine more than 1.600 article abstracts for topics in hundreds of brain regions and then comparing the topics in each brain region with a statistical test to see if the distributions of their coordinates were similar or not.

The biggest difference in this project from what Nielsen et al. did was to use a predetermined list of cognitive terms to data mine the article abstracts with. By using a predetermined list of cognitive terms the topic mining was steered towards topics involving only those terms. The list that was used was obviously not exhaustive, and it was especially lacking in emotional words such as anger, joy, love, hate, sorrow or shame. It could also have been interesting to add a list of diseases to the words that were mined for to see how diseases relate to cognitive and emotional processes in certain brain regions.

The topic mining resulted in more than 2.000 topics in 376 brain regions with over 7.000 topic comparisons. Analysing the results from all of that would be very time-consuming and a far larger task then the time frame of this project allowed. To see if the results were any good the topic comparisons from the posterior cingulate cortex, anterior cingulate cortex and superior temporal sulcus were compared to previous studies of those areas. Even though the topic mining in this project did not always result in the exact topics discussed in the studies about those brain regions some similar topics were found and used as substitutes. The results found in this project showed to be quite similar and sometimes the same as in those previous studies. Taking examples from three brain regions is of course not a proof that the methods of automatic meta-analysis described in this thesis works, but it is a good indicator that it does.

Bibliography

[1] Cognitive paradigm ontology (cogpo), February 2012. URL https://wiki.birncommunity.org/display/NEWBIRNCC/Cognitive+Paradigm+Ontology+(CogPO).

[2] A short history of stereotaxic data volumes at the mni, February 2012. URL http://www.bic.mni.mcgill.ca/~louis/stx_history.html.

[3] Sumsdb (surface management system database) and webcaret online visualization, February 2012. URL http://sumsdb.wustl.edu/sums.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003. URL http://www.crossref.org/jmlr_DOI.html.

[5] Randy L. Buckner, Denise Head, Jamie Parker, Anthony F. Fotenos, Daniel Marcus, John C. Morris, and Abraham Z. Snyder. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, 23(2):724–738, 2004. URL http://www.sciencedirect.com/science/article/B6WNP-4D8VH5W-2/2/f0f85a3eba599128c233f734419c83c9.

[6] G Bush, P Luu, and Mi Posner. Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10827444.

[7] Grit Hein and Robert T Knight. Superior temporal sulcus–It's my area: or is it? *Journal of Cognitive Neuroscience*, 20(12):2125–36, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18457502.

[8] Fox PT. Laird AR, Lancaster JL. Brainmap: The social evolution of a functional neuroimaging database. *Neuroinformatics*, 3:65–78, 2005.

[9] Jack L Lancaster, Diana Tordesillas-Gutiérrez, Michael Martinez, Felipe Salinas, Alan Evans, Karl Zilles, John C Mazziotta, and Peter T Fox. Bias between mni and talairach coordinates analyzed using the icbm-152 brain template. *Human Brain Mapping*, 28(11):1194–1205, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17266101.

[10] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999. ISSN 0028-0836. doi: http://dx.doi.org/10.1038/44565. URL http://dx.doi.org/10.1038/44565.

[11] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000. URL citeseer.ist.psu.edu/lee01algorithms.html.

[12] F.Å. Nielsen, D Balslev, and L.K. Hansen. Mining the posterior cingulate: Segregation between memory and pain components. *NeuroImage*, 27(3): 520–532, 2005. doi: 10.1016/j.neuroimage.2005.04.034.

[13] F. Å. Nielsen. The brede database: a small database for functional neuroimaging. In *NeuroImage*, volume 19. Elsevier, jun 2003. URL http://www.imm.dtu.dk/~fn/Nielsen2003Brede_abstract/Nielsen2003Brede_abstract.html. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19-22, 2003, New York, NY. Available on CD-Rom.

[14] F. Å. Nielsen. The brede wiki: A social neuroinformatics web-service with structured information from neuroscience, jan 2009. URL http://neuro.imm.dtu.dk/wiki/.

[15] Numpy. numpy.finfo — numpy v1.5 manual (draft), February 2012. URL http://docs.scipy.org/doc/numpy-1.5.x/reference/generated/numpy.finfo.html.

[16] J. Ojemann, E. Akbudak, A. Snyder, R. McKinstry, M. Raichle, and T. Conturo. Anatomic localization and quantitative analysis of gradient refocused echo-planar fmri susceptibility artifacts. *Neuroimage*, 6:156–167, 1997.

[17] Peter and Bandettini. Functional mri today. *International Journal of Psychophysiology*, 63(2):138–145, 2007. ISSN 0167-8760. doi: 10.1016/j.ijpsycho.2006.03.016. URL http://www.sciencedirect.com/science/article/pii/S0167876006000985. Cognitive Neuroscience: Contributions from Psychophysiology.

[18] R.A. Poldrack, A Kittur, D Kalar, E Miller, C Seppa, Y Gil, D.S. Parker, F.W. Sabb, and R.M. Bilder. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, 5 (17), 2011. doi: 10.3389/fninf.2011.00017.

[19] E T Rolls. A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, 4(3):161–190, 1990. URL http://www.taylorandfrancis.com/.

[20] J. Talairach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain*. Thieme, New York, 1988.

[21] Princeton University. Princeton University "About WordNet.", February 2012. URL http://wordnet.princeton.edu.

[22] The Pennsylvania State University. Two-sample hotelling's t-square, February 2012. URL http://sites.stat.psu.edu/~ajw13/stat505/fa06/11_2sampHotel/01_2sampHotel.html.

[23] David C Van Essen. Lost in localization–but found with foci?! *NeuroImage*, 48(1):14–17, 2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2756522&tool=pmcentrez&rendertype=abstract.

[24] David C. Van Essen. Personal communication, November 2011.

[25] Jessica B Voytek and Bradley Voytek. brainSCANr, 2010. URL http://www.brainscanr.com/.

[26] Wikipedia. Lemmatisation, February 2012. URL http://en.wikipedia.org/wiki/Lemmatisation.

[27] Wikipedia. Stemming, February 2012. URL http://en.wikipedia.org/wiki/Stemming.

[28] Wikipedia. tf*idf, February 2012. URL http://en.wikipedia.org/wiki/Tf*idf.

[29] Tal Yarkoni, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670, 2011. URL http://dx.doi.org/10.1038/nmeth.1635.