

Data Mining in fMRI Literature:

An Analysis of Bias in fMRI Research on Subcortical Areas
using Natural Language Processing Techniques.

Michiel Boswijk
10332553

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligentie

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor

dhr. G. de Hollander MSc

Co-Supervisor

dhr. dr. I.A. Titov

Institute for Language and Logic
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

June 26th, 2015

Abstract

Research using fMRI has been important in the field of neuroanatomy, and has expanded our knowledge on a large number of brain areas. However, functional neuroimaging research on subcortical brain areas appears to be limited: only a small part of all these areas have been researched. This bias is examined by applying data mining and natural language processing techniques in fMRI studies. The search engine PubMed is used to collect a list of papers mentioning certain subcortical areas. Next, using a brain atlas and a database containing a list of papers with their reported brain activation areas, a comparison is made between the number of times a nucleus is mentioned and whether its presence is indicated by the reported activations in these same papers mentioning them. After that, relations between subcortical areas and cognitive functions are found by applying a word vector technique called word2vec to the text of a set of papers. Even though the results are based on a limited set of data, the results show a bias towards the striatum and show promising relations between nuclei and cognitive concepts. By demonstrating the potential of a meta-analysis, the aim is to encourage researchers to both share their data and to study more subcortical areas using fMRI.

Keywords: *text mining, data mining, natural language processing, subcortex, neuroanatomy, pubmed, entrez programming utilities, medline database, word2vec*

Contents

1	Introduction	4
2	Literature Review	5
3	Research Method	7
3.1	Nuclei Mentions Versus Voxel Reportings	7
3.1.1	Data	7
3.1.2	PubMed Retrieval	8
3.1.3	Probability Calculation	9
3.2	Nucleus Cognitive Function Relations	10
3.2.1	Data	10
3.2.2	Word Vector Model	11
4	Results	12
4.1	Nucleus Mention Probabilities	12
4.2	Nucleus Cognitive Function Results	14
5	Evaluation	18
5.1	PubMed Evaluation	18
5.2	Probability Evaluation	18
5.3	Nucleus Cognitive Function Evaluation	19
6	Conclusion	20
7	Discussion	20
8	Acknowledgements	24
A	Appendix	25

1 Introduction

As the amount of publicly available data grows, data mining offers a way for both the corporate and the academic world to make sense of large amounts of data. In the academic world, relationships between concepts can be confirmed (or at least strengthened) by automatically examining a large set of documents. In this thesis, data mining and natural language processing (NLP) techniques are used in the domain of neuroscience.

The proposal for this project originates from the Birte Forstmann research group. This group does research on the subcortex using fMRI. One aim of this research group is to relate distinct brain areas of the subcortex (also called subcortical nuclei) to cognitive functions (such as memory or perception). A number of these nuclei are indicated in Figure 1¹. Discovering these relations could assist the treatment of patients that experience problems with related cognitive functions. One of the findings of the research group is that a large number of the subcortical nuclei have not been well researched using fMRI. It appears that there is a bias towards certain nuclei in fMRI literature while different ones might be equally relevant (Alkemade et al., 2013).

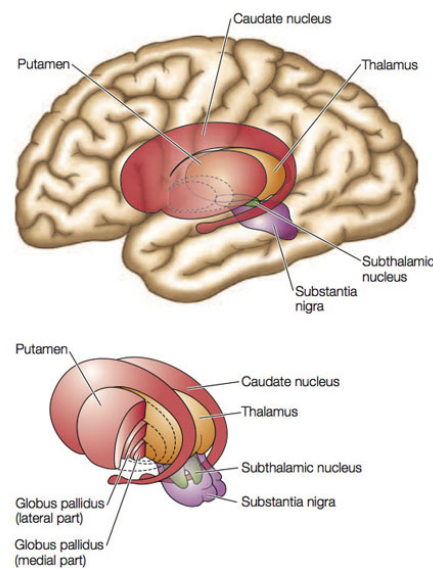


Fig. 1: Human subcortex with indicated nuclei

This thesis has two major goals. The first goal is to test the hypothesis that certain nuclei are ignored in fMRI literature. In the approach to this research, the search engine PubMed (PM) is used, which has access to the MEDLINE database: a database containing over 20 million papers from journals associated with biomedical sciences. Using PM, unique references to papers called PubMed-IDs can be obtained after searching for specific keywords. The search engine allows for different search specifications, such as searching for keywords in the abstract, title or Medical Subject Heading (MeSH terms) of a paper. An application program-

¹ Source: <http://sites.sinauer.com/animalcommunication2e/chapter08.07.html>

ming interface (API) called Entrez Programming Utilities, or eUtilities (see Sayers, 2009), is used to automatically obtain a list of IDs after searching for a nucleus as a keyword. This set of returned paper IDs per nucleus is compared to a set of papers reporting brain activation areas that indicate the presence of said nucleus. The centre of such an activation area is reported in these papers as a voxel (which could be seen as a 3-D coordinate) in some common space of the brain. Consequently, when activation areas are reported that indicate the involvement of a nucleus while this nucleus is not reported, an indication that nuclei are left out of fMRI literature is found.

The second goal of this thesis is to find relations between subcortical nuclei and cognitive functions, which is a more exploratory part of the project. It will be examined whether techniques in artificial intelligence can be valuable for the discovery of these relations. In the approach to this part of the thesis, a vector space model is used to find words that strongly correlate to each other. Words indicating cognitive functions with the strongest correlation value to a nucleus can be obtained this way.

These two goals can be formulated together into one research question:

“Are some subcortical nuclei consistently ignored in neuroimaging studies, even though there is consistent evidence for their involvement in cognitive processes in these same studies?”.

The results are expected to indicate, firstly, a bias towards certain nuclei, and secondly, to indicate strong relations between specific nuclei and cognitive functions. In short, this project can be described as an application of AI-techniques to provide a meta-analysis in the domain of neuroscience. The hope is to encourage researchers to use fMRI to explore more subcortical brain areas and to share their data by demonstrating some of the possibilities that a meta-analysis brings.

2 Literature Review

The problem that led to this paper’s research question is described in Alkemade et al. (2013), where the authors point out the underrepresentation of subcortical areas in fMRI literature. It is reported that approximately seven per cent of the 455 distinct subcortical nuclei, mentioned in the Federative Community on Anatomical Terminology (FCAT), are represented in modern brain atlases. The findings suggest that these areas are far less researched than for instance cortical areas. Possible reasons for this lack of research is that, in comparison to cortical nuclei, different analytical techniques must be applied to subcortical fMRI data, which can be attributed to their deeper location in the brain. It is argued that linking publicly available data from numerous resources could contribute to a broader understanding of the human subcortex. In this thesis the potential bias towards certain subcortical nuclei is examined.

One of the problems that starts to surface when attempting to link fMRI data is the amount of data that is made publicly available. As described in Poldrack et al. (2013), numerous difficulties arise from sharing fMRI data, including varying data formats, large datasets, complex metadata and researcher participation. A framework is presented that allows for the sharing of fMRI data on a large scale, attempting to overcome many of the problems mentioned. If such a project would be widely adopted, data mining and data analysis would become easier and more valuable, since errors are minimised by the use of common data formats. Because the project of this thesis could be described as a meta-analysis, it heavily relies on publicly available fMRI data.

A demonstration of the power of publicly available data is presented in Yarkoni et al. (2011). An automated method is introduced for mapping neural states (voxels) to cognitive states. This is achieved by firstly selecting a set of scientific papers with frequent mentions of cognitive concepts. Secondly, a method is implemented for extracting voxel coordinates from these studies. Thirdly, both forward and reverse inferences between voxels and cognitive concepts are obtained. Finally, a classifier is trained to determine the probability of a new activation area being associated with a certain cognitive concept.

In this thesis, as in Yarkoni et al. (2011), publicly available data is used to find meaningful relations between certain neuroscientific concepts. Specifically, the database created after the second step described above is used for finding relations between voxels being mentioned and corresponding nuclei being mentioned. Instead of finding relations between voxels and cognitive concepts, NLP techniques are applied in an attempt to relate subcortical nuclei to cognitive functions.

The primary NLP-technique that is used in this thesis, is a method developed by Mikolov et al. (2013). The authors propose two unique methods for computing vector representations of words from a dataset. Using these vector representations, word pair relationships can be obtained. An algorithm called word2vec, that implements one of these methods, is used in this thesis for the creation of similarity matrices.

3 Research Method

The approach taken consists of three steps. The first step is to see whether evidence for a bias can be found by comparing the probability of a nucleus being mentioned in the paper, to the probability of it being mentioned given that related voxels are reported. The second step is to relate the nuclei to words relevant in the domain (such as words indicating human functions or diseases).

The scripts² for every part of the approach were written in Python (2.7), specifically in the IPython notebook environment (see Pérez and Granger, 2007). The package used most frequently is the pandas package. McKinney (2012) was used for discovering the pandas functionality.

3.1 Nuclei Mentions Versus Voxel Reportings

3.1.1 Data

To compare papers mentioning nuclei to those mentioning voxels, two databases are used. The first database was made available in Yarkoni et al. (2011), and contains data gathered from the introduced framework Neurosynth. This database contains entries for each paper with different columns of information, such as the paper title, authors, year and voxels. Only the following columns are extracted from this database.

x	y	z	ID
-2	-98	-8	9106283

Tab. 1: Example of data gathered from the Neurosynth database

On this database, a second selection process is performed: only the entries with one common voxel space are extracted in order to create a homogeneous set of data. This voxel space is a standard space of the brain where the voxel coordinates are read from, and is used to eliminate the difference in brain shape and size of different participants in fMRI studies. The MNI space, one of the most used standard spaces for reporting voxels, is used in this selection process. This results in a database containing 280228 entries (indicating the number of voxels), and 7669 unique papers. This dataset is particularly interesting since it links PubMed-IDs to voxels, the advantage being that the IDs can be directly compared to a retrieved set of papers from PM. The selected set of ID-voxel pairs is related to a second database. This second database, which originates from a brain atlas created by the Forstmann research group, contains a large number of voxels accompanied by probabilities of them being part of a certain nucleus. The following table shows one row of this dataset.

² All code is available at https://www.dropbox.com/sh/tbmuexy396a8dg3/AADnmyptBWCNr9o_868QmyB8a?dl=0

Nucleus	x	y	z	P
striatum	5	7	-6	0.619383

Tab. 2: Example of data gathered from Keuken et al. (2014)

Two of these datasets are available. The first one is a set with probabilities for voxels indicating the subcortical nuclei *substantia nigra* (SN), *striatum* (STR), *red nucleus* (RN), *globus pallidus*³ (GPi and GPe) and *subthalamic nucleus* (STN). This dataset originates from a brain atlas introduced in Keuken et al. (2014). The second dataset also contains cortical nuclei *inferior frontal gyrus* (IFG), *primary motor area* (M1), *supplementary motor area* (SMA), and *presupplementary motor area* (Pre-SMA), which originates from a brain atlas presented in Neubert et al. (2015). Using both cortical and subcortical nuclei, these two types can be compared and the results of the latter could be strengthened by obtaining similar results for the former.

3.1.2 PubMed Retrieval

The next step is to retrieve a set of papers mentioning the nucleus. PubMed is queried for each of the nine before-mentioned nuclei in order to retrieve such a set of papers. As expected, the collection of returned paper IDs highly depends on the search term used in PM. In fMRI research, often an abbreviation of a long term is used after this term is introduced. For this reason, the nucleus name as well as the commonly used abbreviations mentioned in the previous section were used as a keyword. Also, when using two terms as keywords, PM will return papers where either of the terms is present. This is prevented by placing quotation marks around two-word terms, which results in the following format used as search-string for each nucleus to query PubMed:

```
("nucleus name" OR "synonyms") ...
... AND "humans"[MeSH Terms] AND fMRI
```

The name of the nucleus, as well as synonyms for this name are searched for in all PubMed fields. The same applies to the term fMRI to select a set of papers that likely contain fMRI reportings (voxels). The term humans is searched for in the MeSH terms since neuroimaging studies on humans report this fact in the Medical Subject Heading, which contains a set of characteristic words of a paper. Using these search strings, a set of paper IDs per nucleus is obtained.

³ This nucleus is often divided into two parts (see Figure 1). However, since many fMRI studies do not allow for the precision to distinguish among the two, the nucleus was considered to be one whole.

3.1.3 Probability Calculation

Next the database, voxel (MNI) probabilities and returned paper IDs are linked to create a database containing a list of paper IDs together with a probability for each nucleus being involved in the voxels of the paper. Additionally, for each nucleus it is reported whether they are mentioned or not, meaning whether they are returned by PM after searching for it. A selection from one paper-ID of this merged database is shown below.

Nucleus	P	ID	Mentioned
striatum	0.454645	10366639	y
substantia nigra	0.000000	10366639	n
subthalamic nucleus	0.000000	10366639	n
globus pallidus	0.000000	10366639	n
red nucleus	0.000000	10366639	n
supplementary motor area	0.000000	10366639	n
inferior frontal gyrus	0.271291	10366639	n
primary motor area	0.000000	10366639	n
presupplementary motor area	0.432430	10366639	n

Tab. 3: Example of merged database

Using this database, now the probability of a nucleus being mentioned versus it being mentioned given it has a probability of being involved (as indicated by the reported voxels) can be calculated using the formula below.

The calculation uses the following variables:

- **NM:** Nucleus mentioned.
- **NV:** Nucleus indicated by voxels.
- **Papers:** The number of unique papers in the used database.
- **Mentioned:** The number of papers in the database that is retrieved when searching for the nucleus.
- **In_Voxels:** The number of papers where the nucleus has a probability greater than zero to be involved as indicated by the voxels.
- **In_Voxels_Mentioned:** The number of papers where the nucleus is both retrieved by PubMed, and has a probability greater than zero for being involved as indicated by the voxels.

$$P(NM) = \frac{\text{Mentioned}}{\text{Papers}} \quad P(NV) = \frac{\text{In_Voxels}}{\text{Papers}} \quad (1)$$

$$P(NM|NV) = \frac{\text{In_Voxels_Mentioned}}{\text{In_Voxels}} \quad (2)$$

3.2 Nucleus Cognitive Function Relations

3.2.1 Data

To find relations between nuclei and cognitive functions, the full text of articles is extracted so that on this text a word vector model can be applied. To extract the articles, a method is implemented based on code by Yarkoni et al. (2011). The way this code works is as follows. The code takes a list of PM-IDs as input. For each ID, first the metadata is retrieved by using the eUtilities function eFetch. From the returned metadata, the digital object identifier (DOI) and journal name are extracted. Next, the full text is extracted by manipulating the URL to make sure it points to the full text article. This URL is built using the journal name and the DOI. Since some journal websites initially only show part of the paper when the correct URL is entered⁴, the address is manipulated once more to attempt obtaining the full text. This can for instance be achieved by substituting the word “abstract” in the URL with the word “full”. The journal in turn checks the privilege of the user to determine whether the full text can be shown. Finally, the website’s source code is extracted to obtain the html. Since this URL-composition is different for each journal, the html extraction currently only works for a selected set of journals: the ones of which both the URL-composition is known, and to which the University’s library is subscribed. Currently full text articles from the following journals can be extracted.

- PLoS ONE
- Human Brain Mapping
- The European Journal of Neuroscience
- Journal of Cognitive Neuroscience
- Frontiers in Human Neuroscience
- Frontier in Neuroscience

After the html is obtained, it is converted into plain text using a Python package called html2text⁵. This package contains functions for removing all html tags from a document, leaving only the text. The resulting text is processed again by removing non-relevant special characters as well as removing nonsense strings (strings that are larger than 20 characters are not likely to be a word). The resulting collection of texts serves as a corpus for the word vector technique discussed in the next section.

⁴ Python is connected with the Google Chrome browser using a ChromeDriver. Source: <https://code.google.com/p/selenium/wiki/ChromeDriver>

⁵ Source: <https://pypi.python.org/pypi/html2text>

3.2.2 Word Vector Model

As discussed in the previous section, the text of 1520 papers has been extracted which now serves as a corpus for the word vector model. The next step is to apply such a model to this corpus, with the goal of finding relations between nuclei and words indicating their function. This NLP-model is a type of neural network language model (NNLM), which builds on the distributional assumption: words that appear in a similar context are semantically related. In these language models, the distributional assumption is incorporated to provide a learning objective: the representation of a word is learned to predict co-occurring words. These co-occurring words in turn provide the specific assumption made in this section: when a nucleus occurs together with a word describing some cognitive function, they are correlated.

The word vector model is a word embedding technique. This means that words from a vocabulary, which in this case is a list of all unique words occurring in the papers, are represented as N-dimensional vectors, where N is commonly based on the corpus or the vocabulary size. Since this model is a type of NNLM, the values of these vectors can serve as the parameters that are trained in the neural network.

Now, bearing the previously mentioned learning objective in mind, words that appear in a similar context will have

similar parameter settings (similar vector values). Consequently, their distance in an N-dimensional space will be short. This distance, measured by the cosine similarity, is the measure used in this project to indicate the correlation between words. Next, when these N-dimensional vectors are mapped to a low dimensional space (2 or 3 dimensions), for instance using principal component analysis (PCA), these results can be visualised. Figure 2⁶ shows a word embedding visualization for a set of words. Instead of showing the vectors, the words are represented as points in a 2-dimensional space.

The specific model applied in this paper is called word2vec, and is introduced in Mikolov et al. (2013). This model is a word embedding technique which allows you to use one of two efficient architectures. In this project, an implementation⁷ from the Gensim package is used in order to incorporate word2vec in the current Python implementation.

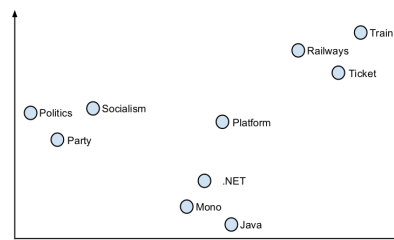


Fig. 2: Human subcortex with indicated nuclei

⁶ Source <http://www.slideshare.net/Azure4Research/bartunov-azure>

⁷ Source: <https://radimrehurek.com/gensim/models/word2vec.html>

4 Results

4.1 Nucleus Mention Probabilities

Firstly, using equation 1, the probability of a nucleus being mentioned, as well as the probability of a nucleus being indicated by the voxels are calculated (third and fourth column) for each of the nuclei. It should be noted that “being mentioned” in this context signifies that the article is returned by PubMed when the nucleus is entered as a keyword, which does not necessarily include all the papers that mention the nucleus once somewhere in the text. The reason for this is explained in Section 5.1. Secondly, using equation 2, the probability of a nucleus being mentioned given that voxels are reported that indicate its presence is calculated (second column). A voxel that indicates the presence of a nucleus is merely the voxel that has a non-zero probability for that nucleus in the atlas database. Thirdly, the first probability is divided by the second to indicate how the mention-probability increases as the associated voxels are reported (column F). Finally, the size measures of the nuclei are extracted from the same atlas database that provided the probabilities (last column). The following table shows these probabilities for all nine nuclei.

Nucleus	P(NM NV)	P(NM)	P(NV)	F	Size (in mm ³)
STR	0.232057	0.124397	0.436041	1.87	14372.7
SN	0.088957	0.033120	0.042509	2.67	307.9
STN	0.028571	0.002738	0.013691	10.43	44.5
GP	0.023947	0.006129	0.157908	3.91	1765.4
RN	0.020253	0.001174	0.051506	17.25	351.5
SMA	0.116949	0.045638	0.153866	2.56	8060.0
IFG	0.074544	0.076933	0.164428	0.97	5812.0
M1	0.033233	0.009388	0.086321	3.54	9112.0
Pre-SMA	0.025199	0.011605	0.294954	2.17	14068.0

Tab. 4: Mentioned and in-voxel probabilities for nine nuclei

Looking at the subcortical areas (indicated by the color blue), the first observation from the results is that there is a substantial difference between the two columns containing probabilities, with the left column containing larger probabilities. This means that these events are not independent, and thereby validates the notion that there is a relationship between the mentioning of a nucleus and whether their activation areas are reported. A second observation is that the probabilities within a column differ greatly, which indicates a preference towards certain nuclei: in this case mainly towards the striatum. A final observation is that even though there does not seem to be a one to one relation between size and the mention-probabilities, size of the nuclei seems to be influential to some extent: the striatum is by far the largest of the nuclei mentioned and also has the largest probability of being mentioned. Because data on the cortical areas is obtained in order to provide some evaluation measure of the subcortical results, these former results are discussed in Section 5.2.

Since the results above are only obtained after looking at voxels that have a non-zero probability for being part of a nucleus, the probabilities are binned according to mentions per nucleus to see the difference between the in-voxel probabilities of the nuclei that are mentioned versus the ones that are not. These results are shown below.

Nucleus	Mentioned	Probability
striatum	n	0.488778
	y	0.554552
substantia nigra	n	0.104446
	y	0.130808
subthalamic nucleus	n	0.042553
	y	0.080374
globus pallidus	n	0.117600
	y	0.211155
red nucleus	n	0.133172
	y	0.227045

Tab. 5: Binned in-voxel probabilities per mentioned and not-mentioned nucleus

For each of the subcortical nuclei, the average probability is taken of the voxels that mention it, versus the ones that do not. In each of the rows of the table, the expected result can be observed: the voxels that indicate the nuclei that are mentioned have a larger average probability than the ones that do not.

Besides binning the probabilities of the voxels into nuclei or mention—bins, $P(NM)$ are calculated for a number of bins to see whether the probability of a nucleus being mentioned actually increases as the probability of the voxel indicating it increases. The following plot is created using a bin size of 10. A regression analysis (least squares) is executed which results in the plotted line through the scattered points. To see whether these results are statistically significant, a Pearson correlation test is executed on this data. The outcome of this test is a correlation of 0.9141 indicating a strong positive correlation, and a p-value of 0.0002 indicating that significant results are obtained.

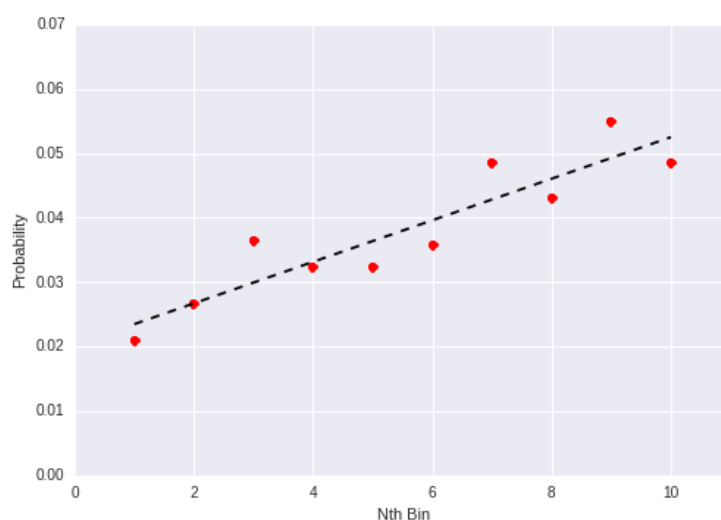


Fig. 3: P(NM) for 10 in-voxel probability bins

These results therefore confirm another expectation: when the in-voxel probability increases, the probability of a nucleus being mentioned increases.

4.2 Nucleus Cognitive Function Results

For relating the nuclei to cognitive functions, word2vec is trained on the text of 1520 extracted papers, using the standard parameter settings (skip-gram architecture, training size of 100 and hierarchical softmax). When the model is trained, first a data driven approach is taken, which is looking at the most similar words for each of the nuclei. Using the function `model.most_similar()`, the top N most similar words in terms of their vector cosine similarity are obtained. The following list of words is obtained when searching for the top 10 most similar words for the nucleus “striatum”.

- | | |
|----------------------------|-----------------------------------|
| 1. ventral_striatum, 0.753 | 6. PFC, 0.735 |
| 2. amygdala, 0.752 | 7. hypothalamus, 0.723 |
| 3. caudate_nucleus, 0.746 | 8. insula, 0.716 |
| 4. dorsal_striatum, 0.741 | 9. is_involved, 0.706 |
| 5. caudate, 0.735 | 10. parahippocampal_gyrus', 0.705 |

These result seem accurate: it is sensible that subareas of the striatum are closely similar to the striatum itself (such as ventral and dorsal striatum or the caudate nucleus). It is also sensible that the striatum, being one of the largest subcortical nuclei, is similar to other important deep-located areas such as the hypothalamus or the parahippocampal gyrus. More interesting is the relation between the striatum and a cortical area like the prefrontal cortex (PFC). Fronto-striatal connections are

important in many cognitive tasks, like researched in Antzoulatos and Miller (2014) and Forstmann et al. (2010).

Even though these results could be valuable for finding inter-nucleus relations, the objective is to relate the nuclei to cognitive functions. To do this an expertise driven approach is used. Experts in the field from the Birte Forstmann research group established a list of words that could be relevant for describing the function of different nuclei (available in appendix A). This list contains not only words describing cognitive function such as vision, taste or motor control but also words describing hormones such as serotonin or dopamine, diseases as well as some nuclei. The initial list was about 80 words long. The function `model.most_similar()` was used to find related words, which expanded the list to about 180 words. Now for each nucleus the most similar words out of this list can be obtained. This resulted in the following correlation matrices, with the correlation values being the cosine similarity values between words.

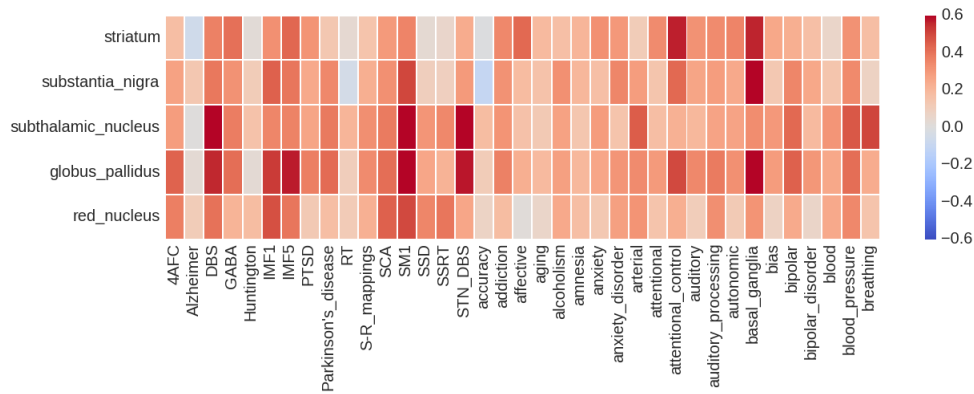


Fig. 4: Correlation matrix relating function to subcortical nuclei (1)

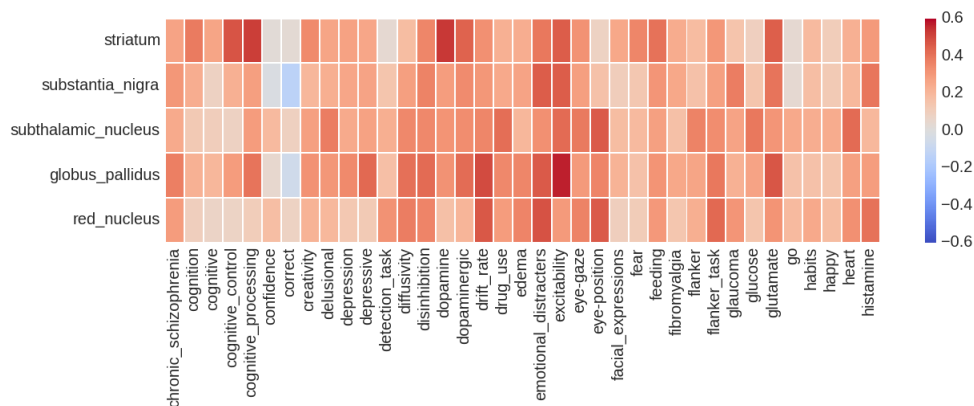


Fig. 5: Correlation matrix relating function to subcortical nuclei (2)

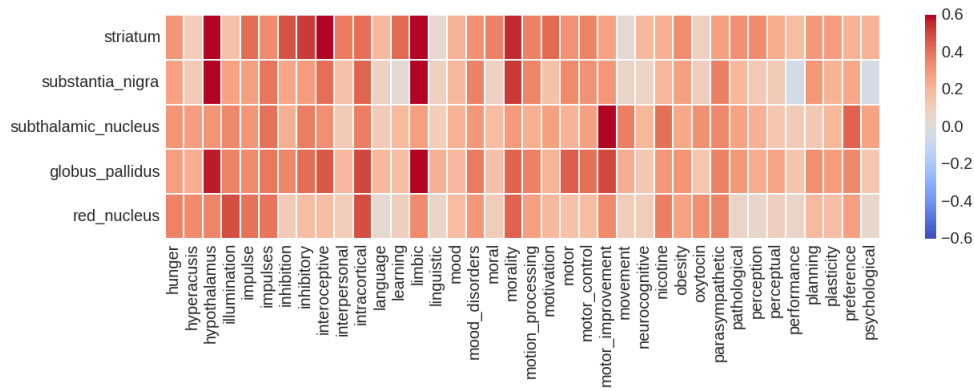


Fig. 6: Correlation matrix relating function to subcortical nuclei (3)

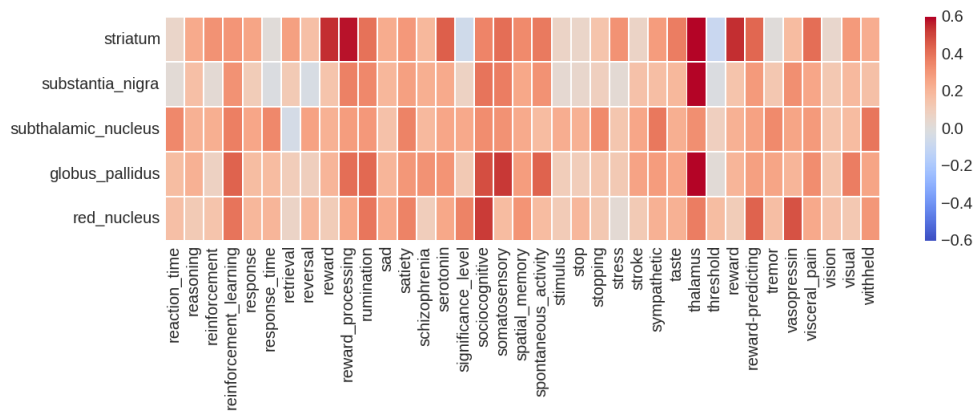


Fig. 7: Correlation matrix relating function to subcortical nuclei (4)

From these results, one can observe words from the list that strongly correlate to each nucleus (red color), and words that have a negative relation (blue color). This representation contains a lot of data and is useful for comparing nuclei, but might not be best for directly viewing the most similar function words per nucleus. A wordcloud is created to provide an immediate overview of the important words per nucleus. The wordcloud for the subthalamic nucleus is shown below.

5 Evaluation

Many of the results should be validated, or are built on assumptions that need to be validated to allow for legitimate conclusions. In this section, first the PubMed search functionality is discussed, then the probability results are compared to results obtained for cortical nuclei, and finally the relations obtained in section 4.2 are validated.

5.1 PubMed Evaluation

As noted in section 4.1, not all papers that mention a certain keyword are returned by PubMed when searching for it. PM does allow for a large number of search specifications. This means that among others, one can search in the author list, title, volume and publisher (for a full list see the PubMed website⁸). However, it does not have access to the full text of papers. It does provide links to the full text, but these links merely redirect to the journal's webpage.

The probability estimates obtained in section 4.1 are obtained by searching all search field available on PM. This means that the set of returned papers after searching for a keyword contains papers in which the keyword plays a relatively large part (as the keyword is reported in one of the searchfields, like abstract, title or MeSH terms). The downside is that it leaves out papers that might have some relevance, namely the set of papers that do mention the nucleus somewhere but not in one of the fields. These papers could still be significant to the study. Unfortunately the full text extraction could not be performed for all papers, as mentioned in Section 3.2.

5.2 Probability Evaluation

In order to validate the findings concerning subcortical areas obtained in Section 4.1, ideally a second database with voxel-paper pairings is available, as well as a subset of these papers mentioning a nucleus (the papers must be searchable). The bias could then be confirmed by obtaining similar results using the new data. Unfortunately, to the best of my knowledge, no such database is available. PM is rather unique in its ability to access a large biomedical database that also has access to the metadata. However, it is possible to compare the subcortical data to results obtained using cortical areas. These results were reported in the pink rows of Table 4.

The results show that, with the exception of the inferior frontal gyrus, the probabilities of a nucleus being mentioned are again larger when voxels are reported that point to them. This confirms the earlier observation that these probabilities are

⁸ <http://www.ncbi.nlm.nih.gov/pubmed/advanced>

not independent from each other. The cortical results also show that there is variation between probabilities in the same column. However (mainly in the $P(NM)$ column) the subcortical probabilities seem to deviate more from each other, which suggests that a bias is more present in the subcortical results. A last observation is that there does not seem to be a direct relation between nucleus size and the probability of it being mentioned for the cortical results.

5.3 Nucleus Cognitive Function Evaluation

To test whether the word2vec method provides valuable relations, an expertise test is created. A wordcloud with the 30 most similar words from the list of relevant concepts is generated for each of the nine nuclei and one for the medial part of the globus pallidus (GPi). Additionally, ten random wordclouds are created from the same list by assigning random correlation values (within the same range) to words from the list. Next, four members of the Birte Forstmann research group that all have extensive knowledge on subcortical nuclei and their functions are asked to select the wordcloud that contains the most words correlated to the nucleus. Three out of the four participants performed above chance level. This was confirmed using a binominal test, which resulted in the following p-values:

While the number of test participants is small, these results provide an indication that meaningful relations have been found. Most of the participants distinguished the ‘real’ wordclouds from the randomly created ones with a p-value that is smaller than 0.05 (indicating a statistical result). Note that these results are based on a set of 1520

Tab. 6: Binominal expertise-test results

Participant	score	p-value
1	8/10	0.011
2	8/10	0.011
3	8/10	0.011
4	6/10	0.170

papers, which does not allow for an extensive set of research for each nucleus (especially when the data would be biased towards certain nuclei). A larger set of extracted papers would likely produce better results, since either relations would be strengthened by the support of multiple evidence or a more complete set of related functions would be reported.

6 Conclusion

In this research, publicly available data was used to examine a bias in fMRI research on the human subcortex. Using an API that connects to search engine PubMed, probabilities describing the likelihood that a nucleus is mentioned were obtained, and compared to the same probabilities given that related voxels are mentioned. Bearing the data limitations in mind, the results of the five examined subcortical nuclei showed a bias towards the striatum, which was far more likely to be mentioned than the globus pallidus, red nucleus, subthalamic nucleus or substantia nigra. When comparing these results to outcomes obtained with data on cortical areas, the former showed more varying probabilities than the latter which suggests that a bias is larger in fMRI literature on the subcortex. The cortical results did not confirm the expectation that larger areas are researched more. However, this finding could be more relevant for areas of the subcortex since these areas are more difficult to research.

Next the full text was extracted from a set of 1520 papers. Using this extracted text as a corpus, a vocabulary was created. Word vectors were created using a type of Neural Network Language Model called word2vec. Consequently the similarity between words in the vocabulary was calculated, which served to measure the correlation between nuclei and a list relevant words. This list of words was created by experts in the field, and contains words describing cognitive function as well as diseases and some nuclei. Correlation matrices as well as wordclouds were generated to visualise these results. An expertise test was created to evaluate these findings, which showed that three out of four participants could distinguish obtained relations from randomly created ones. These nucleus-function connections could assist research in the field of neuroanatomy by pointing out unexpected relations that could be further researched, or by confirming relations that have previously been established.

7 Discussion

Though some promising results have been obtained, there are still many ways in which the research can be improved. As noted in Section 5.2, the probability results depend on the database used, which in this case is a set of 7669 papers with voxel reportings. To calculate the probability of a nucleus being mentioned in a new database containing voxel-paper pairs (and thus to validate the probability results), one would need, besides the database itself, to either have the full text of all papers available (so the text can be searched for nucleus mentions), or a search engine (like PubMed) that has access to the database would have to be available. Unfortunately, not much of this data is publicly accessible. This calls for the sharing of fMRI data (like discussed in Poldrack et al. (2013)) as well the sharing of the full article text. In this project, the full text was extracted from papers by scraping the html from the journal's web page after building the URL, which is a somewhat cumbersome

method made difficult by the denial of access after multiple requests (by the journal website) and redirecting links. Providing full-text access to a user with the privilege of viewing a paper could not only assist a meta-analysis such as this project, but could assist many other scientific disciplines as similar AI-techniques could be applied in different fields of research.

In this project, five subcortical nuclei were used in order to examine a bias in fMRI research. These nuclei were selected since data on these areas was already available and, consequently, because they were relatively easy to delineate in the brain atlas. This means that a bias can already be observed in the selection process of the nuclei: only these areas are well researched as compared to the more than 400 other subcortical nuclei. In future research, the factors leading to this bias could be examined. One of the expectations of this project was that larger areas are more researched since they are straightforwardly easier to research. However, nucleus size is probably not the only factor leading to a bias. Brain location (deeper areas are harder to research), type of scanner and even a trend in research could all lead to a bias.

Finally, even though at this point not all papers allow for automatic text extraction, the full text of papers could be used as a starting point for future research: if the full text is extracted from the limited set of papers that mention the lesser known nuclei, the word vector similarity measure could provide relations that can be examined further (which might lead to more research on the lesser known nuclei). This would assist the ultimate goal: to decrease the knowledge gap in fMRI research on the subcortex by studying a wide variety of subcortical nuclei.

References

- Alkemade, A., Keuken, M. C., and Forstmann, B. U. (2013). A perspective on terra incognita: uncovering the neuroanatomy of the human subcortex. *Frontiers in Neuroanatomy*, 7(40).
- Antzoulatos, E. and Miller, E. (2014). Increases in functional connectivity between prefrontal cortex and striatum during category learning. *Neuron*, 83(1):216 – 225.
- Berney, A., Vingerhoets, F., Perrin, A., Guex, P., Villemure, J.-G., Burkhard, P., Benkelfat, C., and Ghika, J. (2002). Effect on mood of subthalamic dbs for parkinsons disease a consecutive series of 24 patients. *Neurology*, 59(9):1427–1429.
- Forstmann, B. U., Anwender, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., Bogacz, R., and Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, 107(36):15916–15920.
- Keuken, M., Bazin, P.-L., Crown, L., Hootsmans, J., Laufer, A., Müller-Axt, C., Sier, R., van der Putten, E., Schäfer, A., Turner, R., et al. (2014). Quantifying inter-individual anatomical variability in the subcortex using 7t structural mri. *NeuroImage*, 94:40–46.
- McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. ” O’Reilly Media, Inc.”.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Neubert, F.-X., Mars, R. B., Sallet, J., and Rushworth, M. F. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proceedings of the National Academy of Sciences*, 112(20):E2695–E2704.
- Pérez, F. and Granger, B. E. (2007). Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3):21–29. URL: <http://ipython.org/>.
- Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M. P. (2013). Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7.
- Sayers, E. (2009). *The E-utilities In-Depth: Parameters, Syntax and More*. National Center for Biotechnology Information (US).

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–670.

8 Acknowledgements

I would like to thank the Birte Forstmann Research group for granting me the opportunity to work on this project and for providing me with a workspace. During the project, not only have I learned about some of the potential that AI can bring to a field like neuroscience, I have also learned much about fMRI research and the place that such a research group occupies in the academic world. Special thanks to my supervisor Gilles de Hollander for his patience and guidance necessary in this (personally) somewhat novel discipline. I would also like to thank my co-supervisor dr. Ivan Titov for his advice and for shedding his light on the project.

A Appendix

Expertise list of words relating to the function of nuclei.

4AFC	bipolar	glucose
AGRP	bipolar disorder	glutamate
Alzheimer	blood	go
CART	blood oxygen level	gonadotropin
CRH	blood pressure	gonadotropin releasing hormone
DBS	breathing	habits
GABA	chronic schizophrenia	happy
Huntington	circadian rhythm	heart
IMF1	cognition	histamine
IMF5	cognitive	hunger
NPY	cognitive control	hyperacusis
POMC	cognitive processing	hypothalamus
PTSD	confidence	illumination
Parkinson's disease	correct	impulse
RT	creativity	impulses
S-R mappings	deep brain stimulation	inhibition
SCA	delusional	inhibitory
SM1	depression	interoceptive
SSD	depressive	interpersonal
SSRT	detection task	intracortical
STN DBS	diabetes insipidus	language
TRH	diffusivity	learning
accuracy	disinhibition	limbic
action value	dopamine	linguistic
addiction	dopaminergic	menopause
affective	drift rate	mood
aging	drug use	mood disorders
alcoholism	edema	moral
amnesia	emotional distracters	morality
anxiety	excitability	motion processing
anxiety disorder	eye-gaze	motion processing
arterial	eye-position	motivation
attentional	facial expressions	motor
attentional control	fear	motor control
auditory	feeding	motor improvement
auditory processing	fibromyalgia	movement
autonomic	flanker	n-armed bandit
autonomic output	flanker task	neurocognitive
basal ganglia	flushing	neuroendocrinology
bias	glaucoma	neuroendocrine

nicotine	stop task
obesity	stopping
oxytocin	stress
parasympathetic	stroke
pathological	sympathetic
perception	tactile
perceptual	tanocytes
perceptual decision making	taste
performance	thalamus
planning	threshold
plasticity	reward
preference	reward-predicting
psychological	tremor
reaction time	value-based decision making
reasoning	vasopressin
reinforcement	visceral pain
reinforcement learning	vision
response	visual
response caution	withheld
response time	
retrieval	
reversal	
reward	
reward processing	
rumination	
sad	
satiety	
schizophrenia	
serotonin	
sexual behavior	
significance level	
sociocognitive	
somato-visceral	
somatosensory	
spatial memory	
spondylosis	
spontaneous activity	
stimulus	
stimulus value	
stop	