Structured Data Resources for Autism Spectrum Disorder

Heather Logan

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2019

Abstract

The output of published research is growing at an exponential rate, making it difficult for researchers to keep up to date with even a fraction of new findings. Automated curation of biomedical literature has become necessary to deal with this influx of new information and allows researchers to leverage this information for further innovation. Multiple studies show that automated curation achieves best results when developed with respect to some specific domain. This project proposes an approach for the annotation of literature focusing on Autism Spectrum Disorders. Autism Spectrum Disorders have a complex genetic basis, wide variability in symptomology and are largely heterogeneous with a number of co-morbid disorders, making exact characterization a difficult task. The approach proposed here follows a pipeline of biomedical entitiy recognition, identifying autism phenotypes and extraction of relationships between entities, using Natural Language Processing with various open-source tools and resources. The system is evaluated against manually annotated data and shows performance scores comparable to previous systems. Finally, the system is deployed on a corpus of ASD-relevant articles, allowing us to examine the prevalence of ASD-risk genes and phenotypes, identify relationships between them, and observe patterns among the annotated papers.

Acknowledgements

I would like to thank my supervisor Dr. Ian Simpson for his constant help and guidance throughout this project, and my mum and dad for proof reading this report.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

The output of scientific literature is increasing at an exponential rate, with biological and medical science dominating published research in the US and EU [Foundation 2018]. While this surge of new information has great potential to the biomedical research community for enabling further scientific advancements and breakthroughs [Singhal et al. 2016], it can often be under-utilised. Keeping track of this growing literature is virtually impossible for academics, resulting in wasted time and resources, and the oversight of valuable information embedded in underrepresented publications. A structured and comprehensive representation of this data is necessary in providing more direct access to scientific knowledge. Manual curation has been the gold standard method of organising this data, requiring one or more experts to read and annotate each publication. As well as being an expensive and time-consuming task, this method is not equipped to deal with the scope and volume of growing data and does not guarantee perfect accuracy [Elsevier 2018]. There is also the issue of inter-annotator disagreement as a piece of text may be interpreted in different ways [Harmston et al. 2010].

These limitations drive the motivation for text mining in the curation of scientific literature. Text mining is the process of retrieving and extracting information from unstructured text using a number of computational techniques such as Machine Learning (ML) and Natural Language Processing (NLP) [Baker et al. 2017]. Text mining systems have become commonly used tools for biomedical curation and have been used in studies surrounding gene interactions [Lim & Kang 2018], protein interactions [Yu et al. 2018], and modelling of biological processes and diseases [Fluck & Hofmann-Apitius 2013], among many others.

While studies show that text mining systems are effective in accelerating the curation process [Wei et al. 2012], there is still a lot of work to be done to achieve sufficient accuracy so they may be confidently used as a gold standard resource [D. Karp 2016]. Singhal et al. [2016] highlight a prominent issue of curation systems in their ability to generalise to new data, demonstrating that a system designed for text concerning genetic diseases would have significantly different results when used on text concerning

tropical diseases. While improvements in the re-usability of general curation systems is necessary for future progress, many systems achieve good accuracy when designed for the annotation of text within one specific area (e.g. Alzheimer's disease literature [Fluck & Hofmann-Apitius 2013], cancer related literature [Baker et al. 2017]). This is furthermore emphasised by Lévy et al. [2014] who state that best results in curation systems are obtained when developed with respect to some target domain.

One such area where this is necessary is in the research of Autism Spectrum Disorders (ASDs). ASDs are a range of neurodevelopment conditions with a diverse and complex symptomatology and a wide range of comorbidities. The large heterogeneity of ASD poses the challenges of understanding and defining potential subtypes, and identifying causations or influencing factors of the disease [Al-jawahiri & Milne 2017].
The aim of this project is firstly to assess and develop approaches to extract ASD-relevant information from biomedical literature, with an emphasis on identifying ASD-risk genes and phenotypic information. Secondly, to perform the resulting system on a corpora of ASD-related text to develop association data and highlight potential applications in research.

1.2 Summary of Contributions

The main contributions of this project are;

- Collection and preprocessing of a dataset containing ASD-relevant articles from the public database of biomedical literature PubMed.
- Developing a method for extracting biomedical named entities from unstructured text using the MetaMap tool combined with gene lists and a manually curated list of named entities.
- Creation of a dictionary of ASD terms by both parsing of the Autism Spectrum Disorder Phenotype Ontology, and manual labelling. Using this resource to detect ASD-relevant terms and phrases from text using a stemmed text comparison.
- Developing a method for extracting relations between entities in text using a dependency parsing method combined with a rules to filter and combine valid relations.
- Empirical evaluation of each system, in terms of strict and relaxed precision, recall and F1-score, against sets of manually annotated data.
- Analysis of gene and ASD-phenotype frequencies within subsets of articles. Using Term Frequency Inverse Document Frequency to estimate important sources of phenotypic information.
- Analysis of gene-phenotype associations using both term co-occurrence and direct relation extraction from article abstracts.
- Cluster analysis of annotated papers to observe articles related by ASD concepts.

1.3 Report Outline

- Chapter 2 aims to outline the complex characterisation of Autism Spectrum Disorder and its phenotypic expression, detail the rising rate of ASD-focused research and lay out the current stance on the genetic implications on ASD; discussing the work done in this area. We refer to documented approaches to address problems similar to that of this project, and highlight methods which proved as inspiration to the final development of the system, and discuss alternative methods and their suitability for this task.
- Chapter 3 details the steps taken, decisions and adaptations made during the implementation of each module, outlining how they operate together as an Information Extraction pipeline. The materials used in each stage of the implementation are described and justified, and alternative routes are discussed.
- Chapter 4 concerns the empirical evaluation of the resulting system and at various stages in the development process. Each module is evaluated in terms of precision; which is a measure of correct results against all results obtained, recall; a measure of correct results obtained out of possible results, and F1-score; a harmonic mean of precision and recall. We also discuss the reasons for building on each stage of the implementation, common sources of errors in each module, and a general discussion of the results and whether the performance of the system is adequate.
- Chapter 5 describes the analysis of the results produced by the system on a corpus of ASD papers. We investigate the prevalence of ASD-risk genes and specific associated phenotypes, and explore methods to uncover potential relations between genes and phenotypes. Finally, cluster analysis is performed on the annotated papers in an attempt to realise groupings of concepts within the corpus.

Chapter 2

Background

2.1 Autism Spectrum Disorders and Research

Autism Spectrum Disorders (ASDs) encompass a spectrum of neurodevelopmental disorders with common deficits in social functioning and communicative abilities, atypical interests and behaviour patterns [Mcpartland & Volkmar 2012]. Prevalence of the disorder varies significantly among affected individuals in terms of symptom character and severity [Hewitson 2013]. In addition to these hallmark symptoms, the disorder is largely heterogeneous with a nunber of psychiatric and medical comorbidities including Intellectual Disability, Attention Deficit Disorder (ADD/ADHD), anxiety, bipolar disorder, depressive disorders, metabolic issues, gastrointestinal problems and immune deficits [Croen et al. 2015]. A recent study by the Centre for Disease Control and Prevention estimate 1 in 59 children are affected by ASDs [CDC 2018], an increase of over 15% since 2010 estimates of 1 in 68 children [DSM 2010] and a drastic increase since 1997 reports of 7 in 10,000 [Bryson 1997]. This surge in incidence of ASD is largely attributed to improvements in definitions of the disorder and its comorbidities, diagnostic criteria and diagnostic tools leading to increased public awareness and seeking of medical intervention [Zylstra et al. 2014, Elsabbagh et al. 2012].

Driving these medical advances is an influx of research in the area. Graff et al. [2014] evaluated research within three autism and related neurodevelopmental disability journals between 1997-2009 and found that there was an increase of articles published per year in two journals from 18-32 articles per year and 32-171 articles per year over this time frame. The rates in the third journal stayed consistent. In addition, of all articles categorised under the Autism Spectrum Disorder medical subject heading on PubMed from 1976-2019, 86% were published in the last decade.

2.2 The Autism Phenotype

The autism phenotype refers to the collection of characteristics, observations and symptoms that may be exhibited by an individual with ASD. As described, the symptomology of autism is widely heterogeneous with a broad range of comorbidities, symptoms and atypical behaviour. In an effort to facilitate autism research, McCray et al. [2013] constructed the Autism Phenotype Ontology (ASDPTO). An ontology is a hierarchically structured set of concepts and relations pertaining to a particular domain, often accompanied by descriptive metadata for each individual concept class and relation. They are widely used in biological and biomedical research as they enable computational access for a range of applications including data retrieval, integration and modelling [Gkoutos et al. 2015]. The ASDPTO provides a structured vocabulary for phenotype atypicalities found in individuals with ASD, consisting of 283 concepts distributed over three main classes; 'Medical History', 'Personal Traits', and 'Social Competence.' The ASDPTO also provides a link between ASD phenotypes, and other ontologies such as the ASD diagnostic criteria (ASD-DSM) and the Unified Medical Language System (UMLS). Figure 2.1 illustrates the high level categories of the ontology. The 'Medical History' class expands into aspects of an individuals medical background that may have some relation to their ASD phenotype, including co-morbid diseases and disorders, medical symptoms, previous exposures such as environmental conditions, medications or injuries, and an indication of the primary diagnosis. The 'Personal skills' class covers traits which are evaluated as part of the ASD assessment process [McCray et al. 2013] such as cognitive ability; including abstract and analytic thinking, executive function, emotional control, language ability, motor control abilities, and observations of stereotyped, restricted and repetitive behaviours. 'Social Competence' is also used in assessing ASD and covers an deficits in an individuals ability to interact with others, recognition of social norms and cues, predominantly entered around communication signals and customs, and attainment of general life skills within the home and community. This structured representation of a specific domain within the biomedical field may be utilised in numerous ways. The creators of the ASDPTO demonstrate its usefulness in research through its integration with the Autism Consortium database, abstracting the lower level details of ASD diagnostic instruments through its mapping to the ASD-DSM, enabling queries based on simple observations of the phenotype. In Section 2.4, we examine some further ways in which it is applied to information retrieval tasks through text mining.

2.3 Genetic Basis

The notion of a genetic basis for autism is well established and supported by a number of studies; ranging from twin and family studies to genomic sequencing data. However, the extreme genetic heterogeneity of the disease represents a challenge in identifying the causation for ASD and understanding its exact pathophysiology [Yong An & Claudianos 2016]. Twin and family genetic studies demonstrate a strong genetic effect with results estimating 65-95% hereditability in families [Sandin et al. 2017, Tick

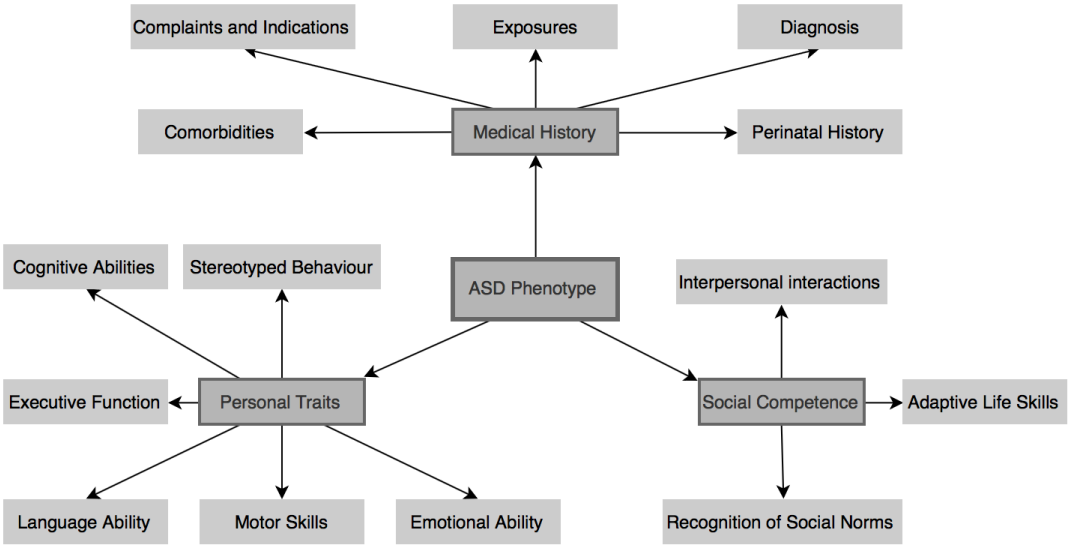

Figure 2.1: Top level structure of the Autism Spectrum Disorder Phenotype Ontology.

et al. 2015]. Genetic scanning continues to validate this genetic basis; with 5-15% of individuals with ASD possessing an identifiable genetic aetiology corresponding to some chromosomal variant or monogenic syndromes; such as Fragile-X or Angelman syndrome [Devlin & Scherer 2012]. The Simons Foundation Autism Research Initiative (SFARI) lists over 800 genes implicated in autism, each with varying confidence [Foundation 2019]. They propose a scoring system to model this confidence in autism-association for each gene. Genes assigned a score from 1 depending on the weight of evidence supporting a genes role in ASD. Genes scored 1 or 2, have a high confidence or are a strong candidate, respectively, and are accompanied by substantial evidence of recurrent and convincing mutations (mutations are likely to be functional) leading to a phenotypic expression. Genes in category 3 or 4, have suggestive or minimal expression, respectively, supporting their role in the disorder. Genes in category 5 are considered hypothesized, with the only supporting evidence originating from model organisms without genetic support in human studies, and genes scored 6 are accompanied by evidence that argues against their role in ASD. The scoring system includes an independent Syndromic category whose members are predisposing to autism as a syndromic disorder (e.g. Fragile X syndrome) and consistently linked to additional characteristics not required for an ASD diagnosis. For example, SHANK3 is assigned a score of 1S as a number of reports identify mutations involving this gene occurring in ASD individuals, and such alterations have been linked to ASD symptoms such as intellectual disability [Zhu et al. 2018] and schizophrenia [de Sena Cortabitarte et al. 2017], and are linked to the frequently co-occurring disorders Phelan-McDermid syndrome [Rubeis et al. 2018] and a Rett syndrome-like phenotype [Hara et al. 2015]. The SFARI gene database currently contains 1053 possible ASD-risk genes, of which 84 are identified as syndromic genes, 25 high confidence genes, and 62 strong candidate genes. In this study, we are particularly interested in the occurrence of these SFARI genes within ASD-related text.

2.4 Related work

With the data explosion in recent years, new attempts to generate a structured representation of information extracted from textual articles are being devised across many fields; from agriculture [Majumdar et al. 2017] to disaster response [Döhling & Leser 2011]. The use of text mining in biomedical domains is particularly prominent due to the rapidly expanding wealth of published biomedical research [M Cohen & Hersh 2005], and is valuable for both researchers and practitioners. The CHAT tool [Baker et al. 2017] uses text mining to retrieve and organise millions of cancer related text and proved it to be useful for cancer research by identifying 'Hallmarks of Cancer' terms and their associations to other entities such as medications and biomarkers within these huge sets of text. Within the autism research sector, Zhang et al. [2016] employ text mining strategies to identify treatments discussed in online autism communities, allowing the creation of a catalogue of treatments and aiding further research on the use of such treatments. Luksic et al. [2016] perform trend analysis over 18,000 ASD-related articles by ontology-based text mining, successfully identifying the main research topics over time and highlighted some potentially under-researched areas.
A text mining approach can be implemented using different strategies, although tend to adopt components from a common framework including data preprocessing, entity recognition and labelling, and relation extraction [Harmston et al. 2010]. Each element of the pipeline may be implemented in various ways, most common strategies for information extraction (IE) include rule-based, dictionary-based and statistical IE. Each approach has both its advantages and setbacks depending on the context in which it is being used. A number of systems applying these approaches were considered against the requirements for the task at hand and the resources available.

Named Entity Recognition

For the task of named entity recognition (NER) in the biomedical domain, a rule-based approach involves consideration of the orthographic features of a word such as particular capitalization or punctuation [Basaldella et al. 2017]. This can be limiting due to grammatical variations however can achieve adequate precision in tasks that follow a specific nomenclature such as mutation identification. MutationFinder [Caporaso et al. 2007] identifies mutations in a standard form (e.g. E6V) or semi-standard form (e.g. Glu 6 to Val) from text using a large set of regular expressions based on a standard mutation nomenclature, achieving a 0.984 precision, 0.819 recall and 0.894 F1-score. Rule based approaches, however, tend not be to robust against new names [Harmston et al. 2010] and would not be effective in cases where a mutation occurrence is expressed in natural language. Use of such rule-based approaches have become rarer in the biomedical domain, however can be useful in conjunction with statistical approaches; as shown in the Bio-NER tool [Soomro et al. 2017] which uses orthographic and contextual features, along with n-grams and affix instances as features for a variety of statistical learners, achieving a final precision of 0.87, recall of 0.866 and F1-score of 0.864.
Statistical approaches often exhibit performance advantages, high precision and im-

pressive generalization results when applied in these systems as they can potentially recognise and label unseen entities [Basaldella et al. 2017]. Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) are popular statistical models for this task. However, to sufficiently train these models, a large manually annotated corpus is required. One example is the Colorado Richly Annotated Full-Text (CRAFT) corpus, manually annotated with concept terms pertaining to nine biomedical ontologies by a team of curators. This provides a valuable resource for the biomedical research community [Bada et al. 2012], allowing the generation of numerous high performance NER models [Basaldella et al. 2017]. The need for a sizable gold standard corpus is a major drawback to this methodology, as manual curation is an incredibly time-consuming task. Furthermore, the need for domain knowledge embedded within the training set means that some gold standard data, such as the CRAFT corpus, is not applicable in all instances. Munkhdalai et al. [2015] demonstrate the importance of domain knowledge within a training set for a biomedical NER task by learning word representations from relevant text for feature selection; achieving a 0.864 F1-score, 0.133 higher than the baseline.

A somewhat more accessible approach to NER is dictionary-based, where entities are compared against one or more dictionaries, databases, or lexicons. An obvious disadvantage to this is that they are unable to recognise concepts out with the dictionaries used, and so are rarely used independently. However, they are often used in conjunction with statistical methods, as demonstrated by Sasaki et al. [2008] who improve the F-score of a statistical NER model from 0.7314 to 0.7872 after utilizing a dictionary of named entities.

A number of systems [Björne et al. 2013, Katona & Farkas 2014] utilise the MetaMap concept annotation tool provided by the Unified Medical Language System (UMLs) which employs syntactic analysis of input text, and look-up against an extensive lexicon containing medical concepts [Aronson & Lang 2010]. MetaMap is explained in more detail in Chapter 3.2. While this is a widely used a state-of-the-art tool, it falls victim to semantic challenges such as term ambiguities and an inability to distinguish between concepts with multiple meanings. The approach used in this task utilises MetaMap while borrowing from some rule and dictionary-based strategies to address these challenges and improve precision with the application of domain knowledge.

Ontology-based Annotation

When performing an entity recognition task within some specific domain, it is necessary to consider the context in which the information is posed. To achieve a domain aware system, the tools should be specialised with respect to some semantic model, such as an ontology [Lévy et al. 2014]. This involves consideration of the vocabulary provided by the ontology while semantically annotating text, and is a popular method for the recognition of specific entities. Lévy et al. [2014] proposes a semantic annotation method with respect to the Gene Regulation Ontology for the recognition of entities and events and their interactions, corresponding to concepts and relations within the ontology. They use this an an initial step in the pipeline for entity and event labelling, then employing Conditional Random Fields for complete extraction of

biological events, achieving 0.61 and 0.50 F1-score on the annotation of events and relations, respectively. Lobo et al. [2017] utilise the Human Phenotype Ontology (HPO) to recognise phenotypic abnormalities within text. They achieve a 0.86 F1-score using machine learning and validation rules based off a gold standard corpora of manually annotated HPO terms. OntoNERdIE [Schäfer 2006] is a system which aims to improve general NER and IE systems through enriching with knowledge extracted from a given ontology. This procedure creates an RDF-structured dictionary of relevant concepts from an ontology in question, and maps to entities during the NER stage of the same structure.
In this task, we are particularly concerned with recognising terms relating to ASD. Thus, the entity recognition system used should be specialised with respect to the autism domain. With access to a specific vocabulary containing such related terms provided by the ASDPTO, ontology based annotation is a sensible approach. While ML methods show good performance, obtaining the resources needed for an ML task, such as a gold standard corpus, would not be feasible for this task. Instead, building a dictionary from the ontology vocabulary, such as Schäfer's [2006] method reasonable, however the final system developed differs from Schäfer's in the structuring of entities. Instead, we use a more simplistic keyword search between text representation of entities.

Relation Extraction

Having extracted named entities from free text, it is necessary to examine their relationships with one another to establish connections between concepts of interest and draw conclusions regarding the subject matter of the text. There is high demand for this in the biological field given the extensive scope for bio-entity interaction, with current RE systems focusing on interactions such as protein-protein [Yu et al. 2018], drug-drug [Björne et al. 2013], gene-phenotype [Li et al. 2018, Khordad & Mercer 2017] and gene-chemical [Lim & Kang 2018]. These systems are dedicated to structure some specific interaction; however many systems employ an open information extraction approach to uncover relations between a general range bio-entities [Elayavilli et al. 2017, Zhou et al. 2014] Such systems typically follow either a rule-based, statistical computational-linguistics based, or a hybrid approach.
Rule-based methods identify relations using inference rules, regular expressions and textual patterns between entities [Xu et al. 2009]. BELMiner [2017] extracts Biological Expression Language (BEL) statements with a rule-based semantic parser using the BEL vocabulary to recognise concepts and biological events such as regulation, binding and transporting events. These events are treated as the primary relation in the sentence, occurring between the concepts identified by the same vocabulary, established by confirming the lexical pattern of the sentence against a small number of rules. "Lack of ability to derive semantic inferences and limitation in the rule sets to map the textual extractions" were reported as limitations of the rule-based system.
Purely statistical-based approaches treat RE as a sequence labelling problem, ignoring the syntactic structure. Such systems do not achieve good performance measures [Harmston et al. 2010], and so are often used in a hybrid approach to apply contextual

and linguistic knowledge.

The use of dependency trees is popular for extraction relations between entities while considering the syntactic structure of the text. Dependency trees are described in detail in 3.5, and an example can be seen in in Figure 3.3. Essentially, a tree representing the semantic structure of a sentence is constructed specifying the relationship between each word, and allows the identification of each entity and the relationship between them. RelEx [2007] implements a dependency parsing system to extract biomedical entities using a set of simple rules and achieves a good precision and recall of 0.78 and 0.79 respectively. Khordad and Mercer [2017] use dependency trees in a hybrid RE method for recognising genotype-phenotype relations, constructing a training set using a number of rules applied to dependency tree representations for a number of sentences, and then extend this to build a self-training algorithm. Both achieve good F1-scores of 0.770 and 0.778, respectively. The construction of a training set in Khordad and Mercer's approach was in itself, a dedicated task and while their self-training process alleviated the burden of manual annotation; statistical or hybrid statistical approaches were not feasible for this task.

The process proposed in the RelEx system is most similar to that used in this task. Due to its reliance on the syntactic structure of text rather than a limited set of pre-specified relations or some relations learned from a large corpus, the dependency parsing method allows for a simple relation extraction without such constraints.

Chapter 3

Methods and Materials

This information extraction system follows a pipeline, outlined in Figure 3.1, divided into Named Entity Recognition, ASD-term recognition and relation extraction. Each module works off of the free text but is able to use the output of the previous module. We make use of open-source NLP tools including tokenization, stemming, stop word lists, part-of-speech (POS) tagging and shallow parsing. Tokenization is the splitting of text into single words, or 'tokens'. Stemming describes a word being stripped of its affixes to its root, and stop words are common words deemed irrelevant to the semantics of a sentence or phrase, such as 'and', 'or', 'be', 'you'. POS tagging is the labelling of words with their appropriate part-of-speech category, including nouns, verbs, adjectives, and pronouns, and shallow parsing is the process of extracting phrases based on the POS tags of their tokens.

3.1 Data Collection

As we did not have access to a specialised corpus, the task of data collection was not to obtain a dataset for which to train a model but instead a corpus containing ASD-relevant text for which to base the implementation and perform analysis upon. Papers were retrieved from the National Centre of Biotechnology Information's (NCBI) PubMed Central (PMC) database of full-text biomedical literature. There were two main criteria for papers of interest; those which contained information regarding the genetics of ASD, and those which contained information regarding the behavioural phenotype of ASD. The paper's full text also had to be openly accessible. Medical Subject Headings (MeSH) is a controlled vocabulary of hierarchical terms and subheadings used for indexing articles within the PubMed/PMC database [Kim et al. 2016]. Articles within the database are associated with a number of MeSH terms, pointing to the subjects relevant to the text. 'Autism Spectrum Disorder' is a MeSH heading itself, with subheadings including complications, genetics, diagnosis, etiology, physiology and psychology. Using the' Autism Spectrum Disorder/genetics' subheading pointed to 3261 papers, of which 1179 were open-access and able to be retrieved. One issue with this approach is that articles will not always be comprehensively labelled with

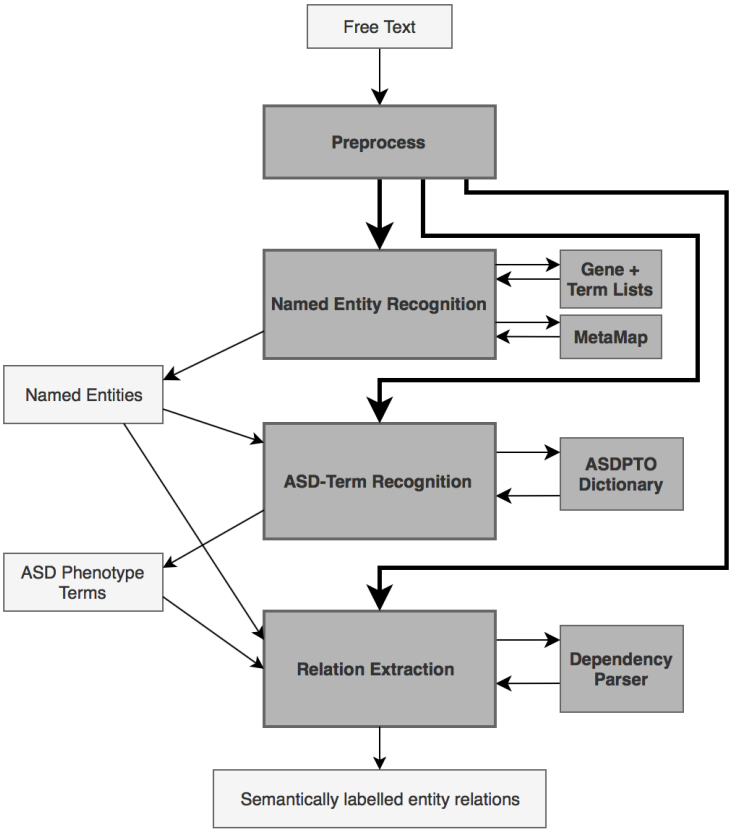

Figure 3.1: General Pipeline of IE System.

relevant MeSH headings, leaving many papers out of our search. On a larger scale data retrieval this should be taken into account, however this approach collected an appropriate number of papers for the purpose of this task. Phenotype papers were collected by combining a query for papers labelled with the 'Autism Spectrum Disorder' MeSH terms, with the constraint that the keywords 'phenotype', social', or 'behaviour' were mentioned in the heading or abstract. This query returned 4998 articles papers however only 1051 were open-access and retrievable. The discrepancy between the number full-text available articles and number of potentially relevant article subjects the collected data to selection bias, which is considered during analysis. The two collections revealed showed 319 overlapping papers, resulting in a final corpus of 1911 papers.

3.2 Preprocessing

Systems often apply a general preprocessing step to their dataset before beginning information extraction, including punctuation cleaning, case conversion and removal of stop words. However, as this task explores a number of modular approaches to IE, it is more suitable to employ different preprocessing methods depending on the pipeline stage. For example, punctuation is necessary for the detection of mutations based on

a mutation nomenclature, and stop words are useful for determining sentence structure at the relation extraction stage. The text underwent general cleaning to remove irrelevant punctuation (e.g. '!', '?', '@'), was converted to lowercase, URLs and links filtered out, and bracketed terms, such as references, removed. Furthermore, acronyms were detected and replaced as to resolve ambiguities that arose in further steps. This involved detecting acronyms inside a bracket which were preceded by terms starting with letters corresponding to the acronym. This was useful for normalisation purposes (e.g. ASD and Autism Spectrum Disorder), and instances where the acronym could not be recognised or was incorrectly recognised for labelling (e.g. MetaMap does not recognise 'SNV' but can recognise 'single nucleotide variant').

3.3 Named Entity Recognition

Named Entity Recognition (NER) is the recognition of some nominal sentence or word from text which can be identified in some set or lexicon [Jimeno-Yepes et al. 2008]. This is a non-trivial task, especially within the biomedical domain, due to the wide variety of biomedical nomenclatures [Harmston et al. 2010]. In this NER implementation, we make use of the MetaMap tool to identify biomedical entities from text, combined with some dictionary-based approaches to improve precision.

MetaMap

MetaMap is a tool providing a link between text and knowledge within the Unified Medical Language System (UMLS) Metathesaurus, an extensive vocabulary of biomedical and health related concepts [Aronson & Lang 2010]. Within the UMLS knowledge source is the Semantic Network, consisting of a range of subject categories, or semantic types, which categorise those concepts within the Metathesaurus [*The UMLS Semantic Network* 2009]. For example, 'disease or syndrome', 'sign or symptom', 'genetic function', 'congenital abnormality'. In summary, MetaMap works by performing lexical/syntactic analysis on input including tokenization, part-of-speech tagging and a look-up against the contained lexicon. Detected phrases and mappings are further improved by identifying term variants and considering other mapping candidates, performing word sense disambiguation to favour mappings that are semantically consistent with surrounding terms [Aronson & Lang 2010]. Figure 3.2 illustrates the recognition of biomedical named entities from an example sentence and labelling with their corresponding semantic types.

Improving Metamap

MetaMap was found to ignore mutation mentions such as c.76A>G or g.74dupA. To account for this, we use a number of regular expressions to search for mutation mentions in the text in their standard or semistandard format, as described by the Human Genome Variation Society's nomenclature [Human Genome Variation Society 2016],

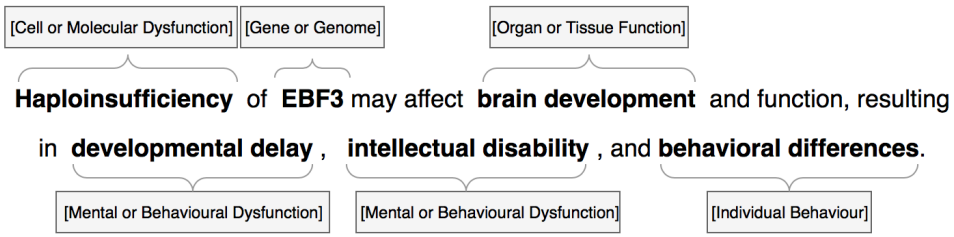

Figure 3.2: An excerpt from Tanka et al. [2017] with named entities labelled with corresponding semantic types.

similar to the method used by Caporaso et. al. [2007]. Furthermore, a number of gene synonyms were ignored by MetaMap. This was improved by performing an initial dictionary look-up against a database of genes with their full name and synonyms, provided by the HUGO gene nomenclature committee [Yates et al. 2017]. It was found that MetaMap returned relatively low precision rates [Chapter 4] and was unable to distinguish between terms with multiple meanings. For example, ASD was labelled as 'Aortopulmonary Septal Defect'. To address this, with the aim of incorporating some domain knowledge into the system; a dictionary was constructed from the 400 most common entities extracted from the corpus, which were labelled and manually cleaned.

Process

We first start by tokenizing data, applying part-of-speech tags, and then perform shallow parsing to obtain noun phrases which are entities. The entities are then checked for occurrences of genes, mutations and terms within the manually labelled dictionary, and assigned a semantic type accordingly. (Genes correspond to 'Gene or Genome', and mutations; 'Cell or Molecular Dysfunction'). The remaining entities are then labelled using MetaMap. The results are then filtered to remove entities which are not related to the biomedical domain (e.g. groups, concepts, objects). This was necessary as the labels assigned to these types of entities was found to decrease precision and such entities are not relevant to our task.

3.4 Identifying Autism Phenotypes

For this task, the aim was to recognise terms and phrases from the text which correspond to ASD phenotypes, or imply the presence of a phenotype. We use the ASDPTO to build a vocabulary of phenotype terms which is then used to compare candidate phrases. The semantic types assigned to entities at the NER stage are also considered in the identification of phenotype terms.

Phenotype Dictionary

As described, the ASDPTO consists of 283 hierarchically structured concepts, each with a label and definition of the corresponding phenotype. It is readily available in OWL (Web Ontology Language) Format [AberOwl 2019], from which we build a list of keywords and phrases that, if found in some term, would indicate the term and phenotype are analogous. The dictionary was built mostly manually, using some shallow parsing of the phenotype title and definition but with the addition of intuitively related phrases, some combination of phrases and symptoms. The concepts of some abnormality, disorder and also personal ability are prominent in the phenotype but expressed in many different variations in real text. To account for this, lists of synonymous terms for each concept were included in the match. For example, 'respiratory impairment', 'respiratory condition', 'breathing abnormality' are recognised as related to the phenotype 'respiratory indications', and 'language skill', 'language proficiency' and 'language comprehension' are recognised as the phenotype 'language ability'.

Matching Phenotypes

First, shallow parsing is performed on the input text to obtain entities consisting of noun phrases, preceded by a verb or verb phrase. For example, 'processing of facial expressions' and 'deficits in social communication'. We then perform a stemmed keyword search against the dictionary. In the first example, the phrase is matched with the dictionary term 'facial process' and mapped to the phenotype 'integrated verbal and non-verbal communication'. In the latter example, the phrase is matched with 'social communication deficit' and mapped to the phenotype 'reciprocal social interaction.' We then perform shallow parsing inside the entity to ensure the correct mapping is assigned. Inner noun phrases are searched against the dictionary. If the inner noun phrase has the same phenotype assignment as the outer entity, we keep the inner entity as this suggests the extended entity is not necessary. Otherwise, we keep the outer entity. This method accounts for phrases such as 'repetitions in behaviour.' The noun phrase here is 'behaviour' which itself is not a phenotype. However, the extended entity 'repetitions in behaviour' shows the presence of the ASD phenotype 'restricted and repetitive behaviour'. In cases where no mapping is detected, we consider the semantic types assigned to entities a the NER stage. However, only a portion are directly relevant to certain phenotypes; for example, 'sign or symptom' to phenotype 'complaints and indications', 'social behaviour' to phenotype 'social competence' and 'congenital abnormality' semantic type to the same phenotype. This is useful when the keyword search is not sufficient. For example; Adderall is a medication commonly taken by children with ASD symptoms. It is not detected as related to any phenotype by the keyword search, however MetaMap detects it as a pharmacologic substance which then corresponds to the 'Medications' class within the 'exposures' group of the phenotype. Where entities have two potential phenotype assignments, we take the lowest phenotype in the hierarchy as final assignment, as this implies a more specific assignment.

3.5 Relation Extraction

The goal of this task is to identify and extract relations between entities in free text. Rather than seeking to identify relations between a specific type of entity, as is the case protein-phosphorylation mining system by Wang et. al. [2014], this approach is concerned with open-information extraction of any named entities that we may be interested in. This is a difficult task due to the unconstrained nature of language; a relationship between two entities may be declared as *"Entity1 is related to Entity2"*, *"A relation between Entity1 and Entity2"*, or *"Relation of Entity1 by Entity2"* [Fundel-Clemens et al. 2007]. We use a method similar to the RelEx system [2007], that uses dependency trees with a number of rules to extract relations.

Dependency Trees

A dependency tree is a structured representation of grammatical relations between words in a sentence [Wang et al. 2014]. A basic example is shown in Figure 3.3. The Stanford Dependency Parser [Manning et al. 2014] provides support for this, using approximately 50 typed dependencies between words.

Sentence tends to follow a structure involving the subject of a sentence, an object or objects which are affected by the subject, and a relation describing this effect. We can determine the role of each word in the sentence with respect to this structure. Using Figure 3.3 as an example, by observing the typed dependencies we can determine 'fragile' as the starting subject of the sentence (indicated by the dependency *'nsubjpass'*), 'expansion' as as the beginning object of the sentence (indicated by the dependency *'nmod'* meaning the nominal modifier of nouns or clausal predicates), and 'caused' as the relation.

The starting subject and object words can be traced to the full subject and object entities using their dependencies (e.g. *compound* entities and modifiers(*'nmod', 'amod'* etc.)), so the subject entity is "Fragile X syndrome" and the object entity is "CGG triplet expansion in FMR1". The Stanford Typed Dependency Manual [2008] provides a comprehensive description of each typed dependency. From this information, we are able to apply rules to extract structured information from free text.

Rule-based extraction

We start by splitting text by sentence. The Stanford dependency parser then constructs a dependency tree from the sentence, specifying the relations and dependency types between words, as well as the POS tag for each word. If the tree contains more than one subject entity; this implies multiple relations within a single sentence and we split the tree into sub-trees, so each relation has a dedicated tree. We then determine the subject and main relation per tree by observing the vertices connected by a subject dependency. In the case of no subject dependency, we take the root vertex as the main relation and the outgoing relations are candidate objects of the main relation. The main relation is

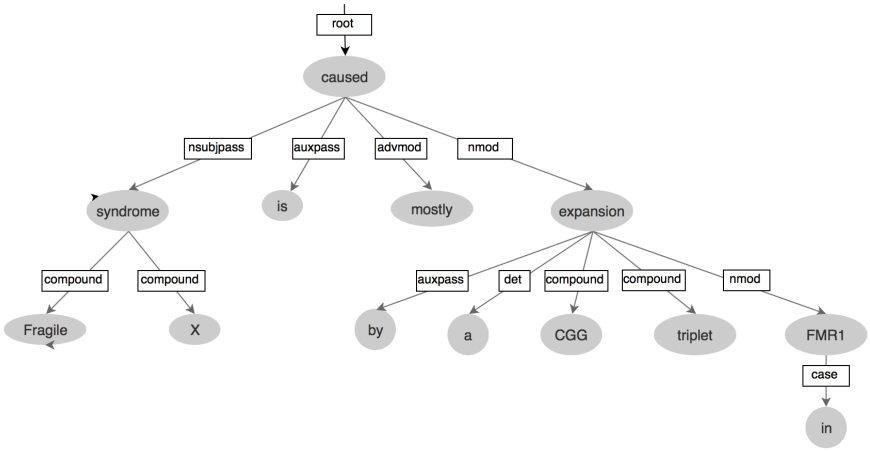

Figure 3.3: An example dependency tree representation of the sentence *"Fragile X syndrome is mostly caused by a CGG triplet expansion in FMR1."*, showing words (elipses), dependencies (arrows) between words, and dependency types (rectangles).

usually the parent vertex of the subject, however to ensure this is correct we check its POS tag and the POS tags of its connected vertices. If the main relation is not a verb but one of its connected vertices is a verb, we set this as the main relation. We take special care in cases where the sentence mentions a correlation between two entities, as the dependency parser was found to often construct an incorrect tree. For this, we set the main relation to the vertex corresponding to the word 'correlation' or 'correlating' etc. An example of this is shown in Figure 3.4. Objects of the sentence are indicated by direct object, modifier or complement dependencies outgoing from the main relation. From this process, we have identified a main relation, a starting subject vertex and a starting object of the sentence. We then trace the path from each of these items to each end point to establish the entire entities. At this stage we break up conjuncted entities as potential relations objects, shown in Figure 3.5. From this process we are given a potential start subject, relation and a list of potential objects. relations are established if there is a valid path between the subject and each object (i.e. the path is not broken by some other relation). Here we can also identify negated relations, which are specified with a negation dependency outgoing from an object vertex or the main relation vertex. A post processing step filters further invalid relations, where some entity does not contain a noun phrase, or occurrence of duplicate relations. We may also identify cases where there is speculation regarding a relation, for example the author states "The gene may affect the phenotype." This is important as the statement does not define a relationship but proposes that the relationship exists [Harmston et al. 2010]. This is expressed in the final relation, so the final relation in the previous example is "may_affect".

This process converts a sentence to the form *relation(entity1, entity2)*. We can then match these to the named entities or phenotypes previously to identify explicit relations between entities of interest. The output given from the examples in figures 3.4 and 3.5 reveal explicit relations between mutations of the SFARI genes- SNAP25 and CEP41- and ASD phenotypes "Over-activity" and "Social Competence", respectively.

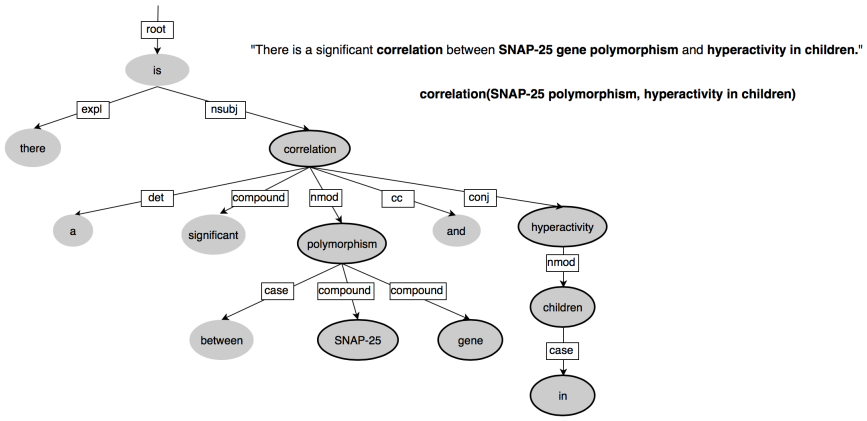

Figure 3.4: Dependency tree and resulting relation showing the main relation being reset to the 'corresponding' vertex.

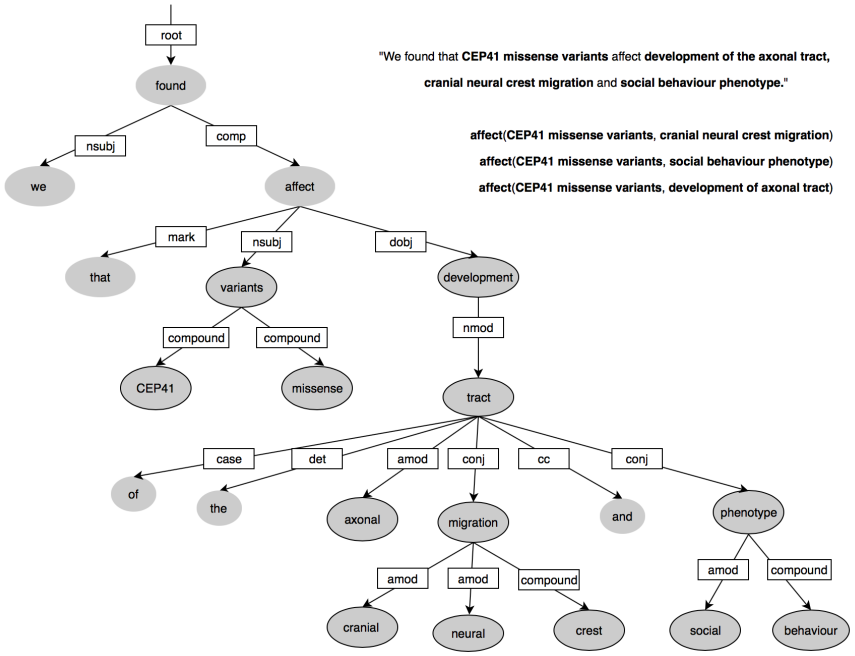

Figure 3.5: Dependency tree and resulting relations showing the splitting of conjunctions and combining with subject.

Chapter 4

Evaluation

To evaluate the performance of the named entity recognition and ASD-term recognition modules, 30 abstracts were randomly selected from the corpus and annotated against the UMLS semantic type list and ASDPTO vocabulary. The NER output was also compared against the open access curation tool PubTator [Wei et al. 2013] which recognises the bio-entities; Disease, Species, Mutation, Chemical and Gene. The relation extraction module was evaluated against a list of 50 sentences containing an explicit relation between two entities, which were manually labelled with corresponding relations. 25 sentences contained gene-phenotype relations and 25 sentences contained a relation between a phenotype and some other entity, such as diseases, exposures or substance. 10 sentences contained an explicit negated relation.
Each module was evaluated in terms of the standard metrics; precision, recall and F-score. Precision is the portion of correct annotations or relations returned by the system that are included in the gold annotation, out of all annotations returned by the system. Recall is the number of correct extractions returned by system out of all items included in the gold annotation. F1-score is the harmonic mean between precision and recall. To account for the recognition of 'almost correct' results; i.e. entities or strings differ slightly from the gold annotation but were still labelled correctly and/or were informative, partial matches were considered in the metrics. The Message Understanding Conference [Chinchor & Sundheim 1993] describe the following sources of errors in comparison of gold-standard annotation and system outputted information:

- Correct (COR): the system output matches the gold annotation.
- Incorrect (INC): the output of a system and the gold annotation do not match.
- Partial (PAR): system and the gold annotation are "similar" but not the same.
- Missing (MIS): a gold annotation is not captured by a system.
- Spurious (SPU): system produces a result which does not exist in the gold annotation.

$$POSSIBLE(POS) = COR + INC + PAR + MIS = TP + FN$$

$$ACTUAL(ACT) = COR + INC + PAR + SPU = TP + FP$$

$$\underset{\text{Precision}}{\text{Strict}} = \frac{COR}{ACT} = \frac{TP}{TP+TN} \qquad \underset{\text{Recall}}{\text{Strict}} = \frac{COR}{POS} = \frac{TP}{TP+FN}$$

$$\underset{\text{Precision}}{\text{Relaxed}} = \frac{COR + 0.5 \times PAR}{ACT} = \frac{TP}{TP+TN} \qquad \underset{\text{Recall}}{\text{Relaxed}} = \frac{COR + 0.5 \times PAR}{POS} = \frac{TP}{TP+FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 4.1: Evaluation metrics where TP = True Positives (Correct and in system output), FP = False Positives (Incorrect and in system output), FN = False Negatives (In gold annotation but missed by system) [Batista n.d.]

4.1 Subsystem Evaluation

Named Entity Recognition

Three iterations of the NER system development were compared. The first is a baseline where extracted noun phrases were labelled using MetaMap only. The second is with the addition of gene lists, mutation regular expression searches, and a dictionary of common ASD-terms found in the corpus. The third is after filtering out 'non bio-entities' in the labelling process. The results are shown in Table 4.1. Each development shows a considerable increase in performance, with a final strict F1-score of 0.635 and relaxed F1-score of 0.725. The low performance of the baseline is largely attributed to textual ambiguities such as multiple interpretations of a phrase or term or multiple possible concepts within a single phrase. The latter issue is discussed more in Section 4.2. The second development aimed to address the former problem by incorporating some domain knowledge from the corpus and strengthen the annotation confidence with additional dictionary look-ups. The second development however still did not achieve adequate performance. On inspection it appeared a large portion of errors resided in the labelling of entities out with the biomedical domain, such as objects and concepts. As they were not relevant to this task, they were filtered out which showed a significant increase in precision and recall.

To measure the performance of the NER system relative to existing tools, 10 abstracts were annotated with PubTator [Wei et al. 2013], a web-based tool for the curation of biomedical papers with respect to various biomedical nomenclatures. PubTator entity recognition is restricted to the specific entities; Diseases and disorders, genes, muta-

	Strict-P	Strict-R	Strict-F1	Relax-P	Relax-R	Relax-F1
Baseline	0.439	0.439	0.439	0.494	0.498	0.495
Second Version	0.555	0.555	0.557	0.616	0.622	0.618
Final Version	0.635	0.636	0.635	0.725	0.700	0.712

Table 4.1: Strict and relaxed evaluation for the Named Entity Recognition system.

tions, chemicals, and species. The entities output by our system were restricted to these types for fair comparison, and the results for each type compared with strict metrics only. The results are shown in Table 4.2. Generally, PubTator achieved good precision with a lower recall than our system. PubTator outperforms our system on the recognition of chemicals, although within the comparison set, only 3 chemicals were mentioned in total. Notably our system outperforms PubTator in the recognition of mutations. PubTator would only recognise mutations in their standard format such as "c.508_511dup", ignoring instances such as "frameshift" and "de novo mutation" which our system is able to recognise. Overall, our system outperforms PubTator on these restricted entity types. Furthermore, our system is adapted to a larger range of entity types, which gives a significant advantage in the context of this task, emphasising the importance of domain-aware tools for domain-specific tasks [Lévy et al. 2014].

	Disease	Gene	Mutation	Species	Chemical	Average
PubTator P	0.920	0.900	0.900	1.000	1.000	0.944
PubTator R	0.550	0.700	0.400	1.000	1.000	0.730
PubTator F1	0.688	0.787	0.553	1.000	1.000	0.805
System P	0.840	1.000	0.850	1.000	1.000	0.930
System R	0.700	0.800	0.860	1.000	0.667	0.805
System F1	0.763	0.936	0.855	1.000	0.800	0.871

Table 4.2: Comparison of NER to PubTator with restricted entity types.

ASD term Recognition

The ASD phenotype recognition module was evaluated first using only noun phrase shallow parsing, as in the NER stage. This method showed a low recall as many phenotype terms are captured within the verbs, such as "behaving", "interacting", "communicating". Implementing the nested shallow parsing approach, discussed in Section 3.4, aimed to address this problem and showed a 23.7% increase in F1-score from 0.573 to 0.709. It was then evaluated when the consideration of semantic types obtained from the NER stage to ; outlined in Table 4.3. The system achieves a final strict F1-score of 0.739 and relaxed F1-score 0.773.

	Strict-P	Strict-R	Strict-F1	Relax-P	Relax-R	Relax-F1
First Version	0.670	0.500	0.573	0.700	0.550	0.616
Second Version	0.740	0.680	0.709	0.800	0.730	0.763
Final Version	0.750	0.730	0.739	0.810	0.740	0.773

Table 4.3: Strict and relaxed evaluation for the ASD Phenotype Recognition system.

Relation Extraction

The final RE module was evaluated independently, also in terms of strict and relaxed precision, recall and F1-score. A result was considered exactly correct if the relation term and both arguments exactly matched the golden annotation, and partially correct if the aspects of the golden relation were captured in the output.

For example; the result of RE on the sentence; *"Novel Causative variants in DYRK1A, KARS, and KAT6A are associated with intellectual disability"*, is;

associated('Novel Causative variants in DYRK1A', 'intellectual disability')
associated('KARS', 'intellectual disability')
associated('KAT6A', 'intellectual disability')

This is partially correct, as although the phrase *"novel causative variants*" should be applied to the genes "KARS" and "KAR6A", the output captures the majority of the correct relation. The final system achieved final strict and relaxed F1-scores pf 0.495 and 0.570 respectively, shown in Table 4.4. On analysis of results, there were no clear differences in the systems ability to recognise gene-phenotype relations and phenotype-other entity relations, rather it depended on the structure of the sentence. The main sources of errors are discussed further in Section 4.2. 6 out of the 10 negated sentence were identified, the rest of which appeared to be a result of an incorrect dependency tree or an incorrect relation result.

	Strict-P	Strict-R	Strict-F1	Relax-P	Relax-R	Relax-F1
Final RE System	0.540	0.450	0.495	0.620	0.570	0.597

Table 4.4: Strict and relaxed evaluation for the Relation Extraction system.

4.2 Sources of Error

The NER module and RE module both make use of publicly available tools and so the correctness of the results heavily depend on the quality of these tools. Although steps to resolve ambiguities such as acronym replacement and lookup against a manually annotated set of common entities, MetaMap's inability to distinguish between ambiguities not annotated in the set was a prominent cause of false positive results. For example, "WT1" may refer to the WT1 gene, or the WT1 nephroblastoma. MetaMap

would not be able to detect which context the term is being used to annotate this correctly. Furthermore, the extraction of entities was restricted to noun phrases only which was a main source of both false positive and false negative results. As described in Section 3.4, the ASD-term extraction used nested shallow parsing to identify concepts that span out with a noun phrase. This was not applied at the NER stage as MetaMap's slow performance and large scope of semantic types meant this method was not practical for this application; entity labelling would take twice as long at best, and the large number of possible types would likely give inconsistent results. The main cause of false positive named entities found was due to MetaMap's recognition of multiple terms within a single entity. For example "human robo1" is a single entity referring to a gene, and should have the semantic type 'Gene or Genome'. However, MetaMap recognises the semantic types for "human" and "robo1" separately, and therefore proposes the candidate semantic types 'Human' or 'Gene or Genome'. As a result, this particular entity is misclassified as type 'Human'.

A main weakness of the ASD-term recognition method is its limited coverage of possible terms and phrases due to the dictionary-based approach. For this reason, it would not perform well in cases where a phenotype is described in terms not included in the dictionary, resulting in the majority of the false negatives found during evaluation. Considering the MetaMap semantic types when labelling with ASD phenotypes alleviated this problem slightly, however there was still the problem of restricting the bounds of the entity to noun phrases and verb-noun phrases. There also exists ambiguity between phrases and a potential phenotype assignment. For example; the term "anxiety" on its own may refer to the 'anxiety disorder' phenotype within the 'comorbidities' class , or may refer to the 'social anxiety' phenotype within the 'social competence' class.

The relation extraction system returned a high false positive rate. On further inspection, it appeared the main reasons for this were due to incorrect combinations of arguments when conjunctions were involved in the sentence, and an incorrect identification of the main relation from the tree. The first issue was mainly due to the filtering stage of the relation construction stage not being strict on arguments which were not directly connected in the sentence. Another reason was the failure to split up complex sentences containing multiple relations or nested relations; where an argument in the result would contain a large section of text. The second issue arose when the input sentence had a root relation which was not the correct relation of interest. For example; "Our study shows gene had effect on phenotype." The main relation is incorrectly identified as the root of the sentence; "shows", which leads to the subject identified as "our study", and the incorrect result; *shows(our study, gene had an effect on phenotype)*, which ignores the correct information in the sentence. As relations are extracted on sentence level, anaphors(e.g. 'that', 'it', 'who') referring to some previous entity from another sentence, were found to be a source of false negatives. A number of false negatives were also caused by incorrect POS tags, causing relationships between entities falsely identified as verb phrases to be filtered out in a post-processing stage.

4.3 Evaluation overview

The system shows various weaknesses, partly stemming from semantic ambiguity, a restricted scope for identifying entities, and common linguistic challenges. Although we expect a ML method to outperform each module, the proposed system achieves fair precision and recall using available tools with the absence of a gold standard dataset, especially considering that manual annotation accuracy itself is usually around 90% [Elsevier 2018]. While a perfectly accurate annotation system is likely not achievable, there is much room for improvement for each module. Taking into account these disadvantages, this system demonstrates an alternative method for information extraction when certain resources are not available, and at times is successful in highlighting important sources of information such as gene-phenotype interactions. This suggests it may be useful when combined with statistical methods, similar to that demonstrated by Khordad and Maryam [2017] and Basaldella et al. [2017].

Chapter 5

Corpus Analysis

Applying the text mining system to the corpus of papers may allow us to gain an understanding of the content of specific papers, perhaps observe the quantity of research revolving around specific genes or phenotypes and highlight specific knowledge of interest such as gene-phenotype interactions and clustered topics in autism research. We look both at the full corpus, and within papers under the Autism Genetics MeSH heading, referred to as 'gene papers', and the papers retrieved considered 'phenotype papers', detailed in Section 3.1.

5.1 General

5.1.1 Gene Prevalence

Genes were counted combining counts from their approved symbol, synonyms and full names, however synonyms were filtered where possible to remove those which were also English words, such as MICE and CAGE. Of 1054 genes in the SFARI database, 794 were mentioned in the full corpus, with 694 in gene papers alone and 462 in phenotype papers alone. On average, a gene paper referenced 6 unique genes; while an average phenotype paper referenced 2 unique genes. Considering the score for each gene (described in Section 2.3), the degree of its ASD-risk is generally reflected in its frequency of document mentions. All 25 high confidence genes and 93% of strong candidate ASD genes are referred to in the full corpus, and generally, representation decreases with lower risk. Figure 5.1 shows the percentage of genes in each category which are mentioned at least once in the full corpus, gene papers only, and phenotype papers. Gene papers generally mention a larger portion of genes in each category than phenotype papers, with the exception of 'strong-candidate' risk genes. Looking at the mean number of document mentions for genes in each category, (Figure 5.2) it is evident that the higher risk genes are discussed more in each collection of papers, and more so in gene papers than phenotype papers.

Figure 5.3 shows the distribution of paper mentions for each gene in categories 1-6, and Figure 5.4 shows the distribution of paper mentions for all syndromic genes.

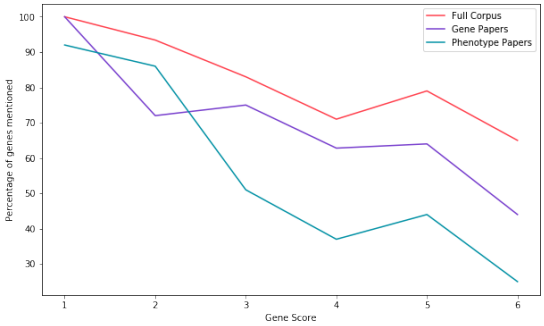

Figure 5.1: Percentage of genes, per category, represented in each collection.

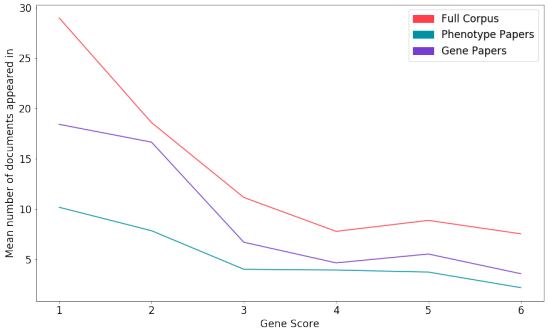

Figure 5.2: Mean number of documents mentioning genes in each category.

The distribution of paper mentions seems to follow a general trend, with a number of significantly higher mentioned genes per category.

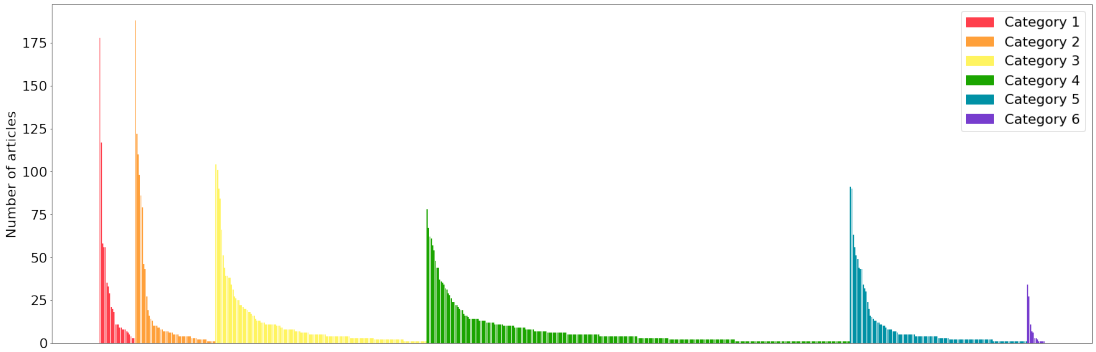

Figure 5.3: Document mentions per gene in categories 1-6.

Looking at the highest mentioned genes in the corpus, most are either high risk, strong candidates, or syndromic (Table 5.1). In particular; FMR1, MECP2, SHANK3, TSC1 and PTEN are syndromic genes strongly associated with ASD co-morbid disorders Fragile-X syndrome, epilepsy, Phelan-McDermid syndrome, Tuberous Sclerosis, and macrocephaly, respectively. As co-morbid disorders are of particular interest in ASD research, it is not surprising that these genes are prominent. The only gene included here outside of these three categories is SHANK1 which currently only has suggestive evidence of implication in ASD, although its family members SHANK2 and SHANK3 are implicated in ASD.

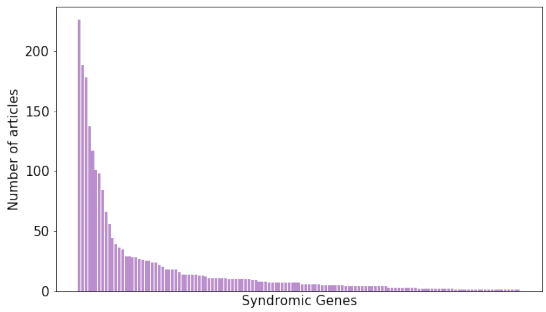

Figure 5.4: Document mentions for syndromic genes.

Gene	Score	Articles Mentioned
FMR1	S	226
MECP2	2S	188
SHANK3	1S	178
TSC1	S	137
SHANK2	2	122
PTEN	1S	117
TRIO	2	110
SHANK1	3	104
MTOR	3S	101
CNTNAP2	2S	98

Table 5.1: Genes mentioned with the highest document frequency in the full corpus.

5.1.2 Phenotype Prevalence

For this study, the semantic types were not considered in the assigning of ASD annotations as labelling entities using MetaMap was too slow to annotate all full text papers in the time given. Instead, annotations are based on the dictionary method alone, described in Section 3.4.

In total, there are 31855 phenotype annotations in the full corpus, representing 256 of the 284 ASDPTO phenotypes. The highest annotated phenotypes are 'learning disorders', 'reciprocal social interaction', 'cognitive ability', and 'stereotyped, restricted and repetitive behaviour', annotated in 75%, 65%, 63% and 47% of papers, respectively. These most common annotations describe general characterisations of ASD and are often mentioned within general statements or ones which summarise the disorder as an introduction to the text; e.g. "ASD is characterised by developmental delay and social deficits". As such, these annotations do not always provide a representation of what the paper is about, or the actual distribution of topics in our corpus.

To normalise this, phenotype prevalence was considered by the number of documents where each phenotype is a prominent topic or concept. This was done using Term Frequency-Inverse Document Frequency (TFIDF).
TFIDF (Figure 5.5) is a measure of importance of a word or term in a document based

its frequency in the document compared to its frequency in a collection of documents. It penalises very common words and considers rarer words as more informative.

$$TFIDF_{t,d} = (tf_{t,d}) \times \log_2 \frac{N}{df_t}$$

where;
$tf_{t,d}$ = frequency of term t in document d
df_t = number of documents term t appears in
N = number of documents in collections

Figure 5.5: Term Frequency-Inverse Document Frequency (TFIDF) Formula

The annotated versions of each article were indexed and the TFIDF against each document was computed. A score of $maxTFIDF \times 0.6$ was found by trial and error, which was a cutoff for a term's associated 'informative' papers. For example; the 'learning disorders' annotation appears in 1444 articles, but is only a prominent topic in 632 of these. Conversely, 'fine motor skills' is only annotated in 75 papers but is considered important in all of these. The distribution of phenotypes by higher class is similar between the total document frequency and TFIDF normalized counts (Figure 5.6), however this metric gives a better understanding of the importance of each term in our corpus.

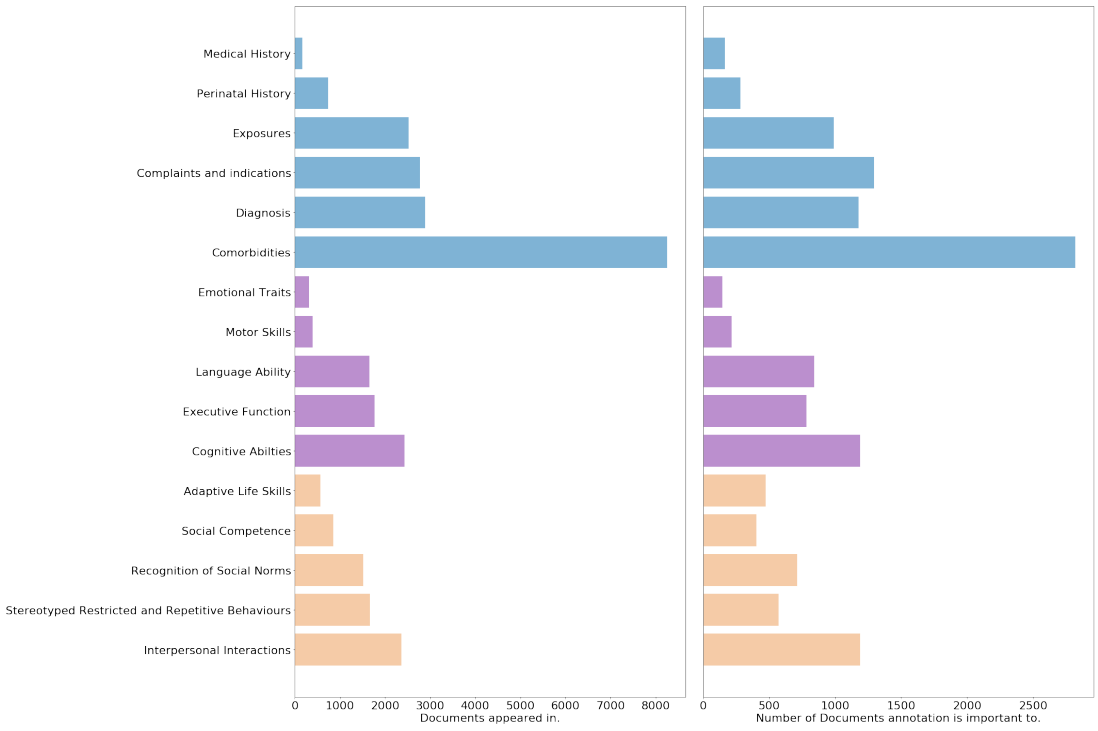

Figure 5.6: Document frequency of phenotype annotations (left) and TFIDF-normalised counts (right), grouped by higher level class. y-axis is same, x-axis is different.

As we see, annotations under the 'comorbidities' class are significantly higher than the

rest. This is not surprising as co-morbid diseases and disorders are of particular interest in ASD research. Looking at the distribution of specific co-morbid disorder annotations (Figure 6.1), mental and learning disorders are particularly prominent, specifically anxiety and depressive disorders. Nervous system diseases the second most prominent phenotype grouping. The distribution of the other types of comorbidities varies between 0-50 documents. The cardiovascular, bacterial and urogenital diseases do not have associated specific comorbidities within the phenotype ontology and so the general disorders are annotated, whereas the other co-morbid disorder types have specific diseases associated and so have a wider scope for annotations.

We may then compare annotation distribution between gene papers and phenotype papers. Gene papers on average contain 10 unique annotations per paper, whereas phenotype papers contain around 16. The higher level distribution of annotations (Figure 5.7) is fairly representative of the overall count; again, comorbidities are a dominating topic of interest in both cases. Both collections have similar distributions, although we can note a significant increase in personal trait and social competence related annotations in the phenotype papers. This is not surprising as we could expect gene papers to be more clinically or medically focused, and phenotype papers to be interested in the behavioural or social aspects of ASD.

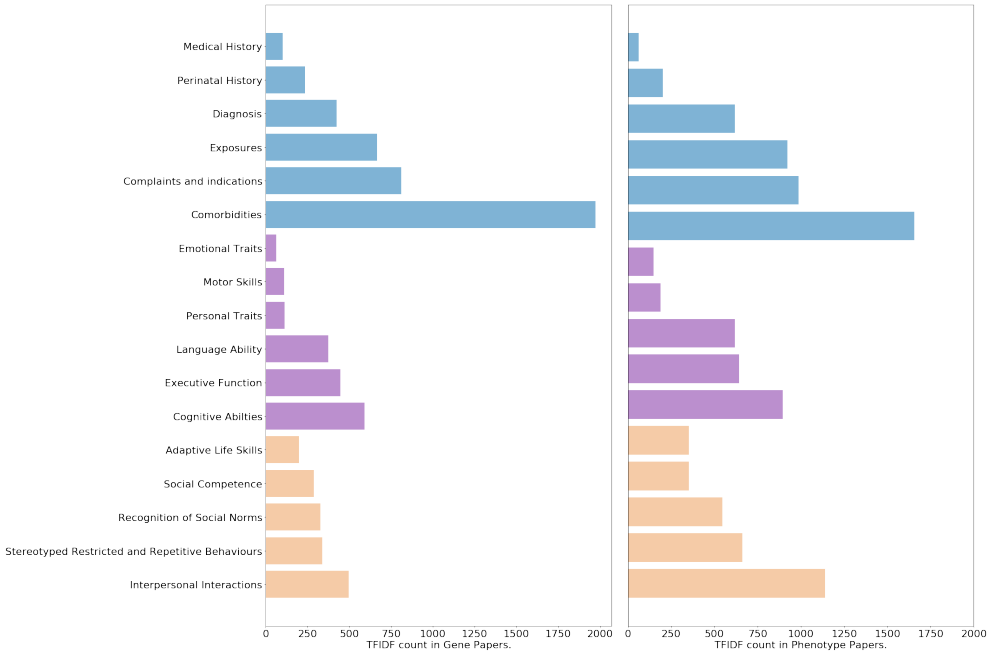

Figure 5.7: TFIDF counts of gene papers (left) and phenotype papers (right).

Emotional traits, such as mood, affect and self concept, are the least annotated phenotype in both collections. We could assume this is an under-researched area; however this may be a result of an annotation issue, as anxiety and depression are considered both emotional traits and co-morbid disorders, resulting in the emotional trait phenotype appearing underrepresented in our papers. This poses the problem of conflicting annotations within the ontology, which may be addressed by considering the context of neighbouring terms using an ML method.

We may use this metric to gain an understanding of the areas research being conducted within ASD research. From our results, it appears co-morbid diseases is very highly researched among genetic and phenotype papers, followed by symptoms and exposures. We can see a difference in the level of research into social aspects of ASD in the genetic research and phenotype research. We have the the issue of selection bias within this small subset of papers, and so this is unlikely to be accurately representative of real research being continued today, however will be applicable to a larger data set.

5.2 Gene-Phenotype Interactions

Of particular interest in ASD research is the relationship between genes and ASD phenotypes. Given the huge heterogeneity of the disorder, both genetically and among symptoms and phenotypes, understanding exact associations is a difficult task. The annotated corpus was examined so that we may observe trends in gene and phenotype mentions amongst the text and form some conclusions about their relation to one another.

5.2.1 Co-occurrence

We may estimate that a gene and phenotype have some relation to one another if they frequently appear together in text. To examine this, the co-occurrence of each gene against each phenotype annotation in the corpus was taken and gene-phenotype pairs which had a significantly high co-occurrence were selected. Certain phenotypes were found frequently with multiple genes and so were grouped. Figure 6.1 shows the phenotypes with the genes they occurred with most frequently, in order of highest co-occurrence.

From the results, we see syndromic comorbidities with high co-occurrence to their causative genes, including; FMR1- Fragile X disorder, MECP2-Rett Syndrome and UB3A-Angelman syndrome. Epilepsy, like ASD, is largely heterozygous and shares many phenotypes with ASD. Epilepsy annotations were found to occur frequently with 10 particular genes, all of which are considered epilepsy risk genes [Wang et al. 2017]. There are numerous pairs found which are have some known association; FOXP1 and FOXP2 mutations have been repeatedly implicated in language disorders [Lai et al. 2001, Bacon & A Rappold 2012], BDNF is a regulator of synaptic plasticity mechanisms underlying learning and memory in the CNS [Cunha et al. 2010], dysfunctions of the MTOR signalling pathway in the brain is associated with a number of neurologic disorders [Lipton & Sahin 2014], and EGR2 is a transcription factor for myelin formation and maintenance, thus alterations can impact motor development [Yiu & Baets 2015].

Despite these confirmed associations identified this way, genes and phenotypes may occur often by coincidence; the FAT1 has a high co-occurrence with Down syndrome although the two do not appear to be associated. Furthermore, this would not identify negative associations, where relationships have been disproven or argued against.

5.2.2 Gene-Phenotype Extraction

While the above method can be informative when considering gene-phenotype relations in text alone, it is not definite and due to the issues described; it is necessary to examine the text for explicit relationships. To compare this method, the relation extraction module was applied to the abstracts of the gene paper subset of the corpus. The comparison was restricted to the article abstracts due to the slow performance of the module, as well as the noise within the full text (e.g. tables, charts, references) which caused a high false positive rate on testing. After filtering results from 1051 abstracts, we found only 108 relations between 48 genes and 35 phenotypes. The sparsity of the results can be attributed both to the relatively low recall of the system, and the fact that relations are often implied explicitly within the text, or out with a single sentence; which our system is not equipped for.

Figure 6.2 shows a network of these relations. The edges are labelled with the relation term extracted from the text. While sparse, the results supported a number of the results estimated previously; including GARB3-epilepsy, MAGEL2-Prader Willi syndrome, SLC9A9-ADHD and FOXP1 and FOXP2-Language Ability. Note that relations may originate from sentences such as *"Mice with GENE mutation exhibit PHENOTYPE"*, so whether this constitutes a definite relation can be speculative, however will usually indicate an association.

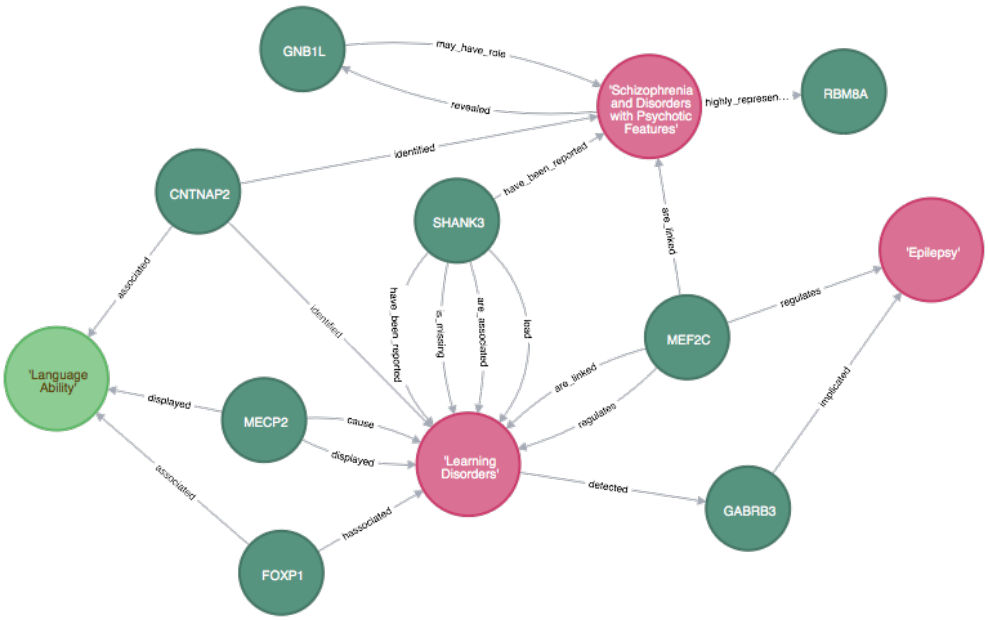

Figure 5.8: A section of interconnected gene (dark green nodes) to phenotype (comorbidities: pink nodes, language ability: light green node) relations extracted from gene paper abstracts.

The system is able to identify associations between other known linked pairs; including IL-6 and Perinatal History; IL-6 has been the subject of investigation for prenatal exposures and has been associated with adverse neurological outcome in preterm infants such as brain injury and autistic symptoms [Rasmussen et al. 2019]. This result shows

a direct impact between a gene and physiological phenotype which may cause or influence the disorder. We also see associations between genes and social or behavioural phenotypes exhibited by ASD, or co-morbid disorders; for example, both MEF2C and CNTNAP2 are linked to learning disorders and schizophrenia, and both SHANK3 and FOXP1 are linked to anxiety disorder and language ability, among others (Figure 5.8). Each is useful in observing mechanisms within ASD genetics, overlapping symptoms and phenotypes displayed by some altered gene or genes, and the genetic similarities between ASD and co-morbid disorders.

5.3 Cluster Analysis

Given the extensive volume of literature available to researchers, it would be useful to organize these into groups of articles related or relevant to one another. Clustering is an unsupervised method of discovering natural groupings amongst a dataset based on similarity and patterns in the content. Clustering is performed on the gene and phenotypes annotations derived from each article, making the assumption that relevant documents contain similar annotations. In clustering the annotations rather than the raw text, the documents are grouped with respect to their genetic and phenotypic information. This will result in some lost context of other subject matter, but we primarily aim to explore the effectiveness of the annotated representation in grouping relevant documents.

In this task, we use K-Means and hierarchical clustering to establish suitable groups. The text is preprocessed by tokenizing and converting to lowercase. Further preprocessing is not necessary as the annotations should be standardized.
A term-by-document matrix is created using TFIDF (Figure 5.5) representing the relative importance of every term in the whole corpus to each document. From this matrix, the distances between each document is computed by $1 - cosine\ similarity$, where cosine similarity is a measure of similarity between two vectors (Figure 5.9).

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Figure 5.9: Cosine Similarity formula, where A and B are document vectors.

Hierarchical clustering groups objects based on their distance in a top-down (divisive) or bottom-up (agglomerative) method resulting in a tree like structure of major and subclusters. We use the agglomerative algorithm which initially considers each document as a single cluster, then consecutively groups similar clusters until all documents are grouped as a single cluster [Jayanthi & Kavi Priya 2018]. K-Means clustering is a partitioning algorithm which defines k number of arbitrary centroids for each cluster and assigns each document to the cluster of its nearest centroid. The centroids are repositioned at the centre of its members, and documents are reassigned recursively until results converge [Ravi & Sundarambal 2018]. Both are used here as K-Means requires a known number of k, which can be estimated from a dendogram constructed by the hierarchical method (example Figure 5.10).

To begin, the full corpus was clustered hierarchically, revealing 3 major clusters and 6 significant clusters at a lower level (Figure 5.10). Performing K-Means with k=3 showed a distinct segregation (Figure: 5.11), however clusters at k=6 were largely overlapping (Figure 5.12) and suggests these groupings are indistinct. We then focus on the three main groups discovered.

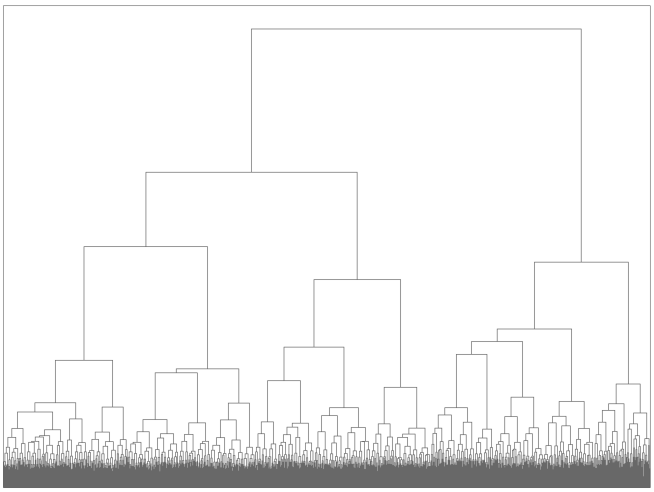

Figure 5.10: Dendogram structure from hierarchical clustering on full corpus.

Figure 5.11: K-Means clustering on the full corpus with k=3.

Looking at the articles contained in each cluster, we see cluster 0 (green) is composed of 87% gene papers, cluster 1 (pink) is composed of 34% gene, 45% phenotype and 20% overlapping papers, and cluster 2 (purple) is composed of 85% phenotype papers. Looking at the most frequently surfacing concepts in each cluster (Figure 5.2), we can observe patterns among concepts within the papers.

Cluster 1 appears to prominently feature syndromic co-morbid diseases including Fragile-X syndrome, Rett syndrome, tuberous sclerosis and epilepsy. Concepts association with the characterization of these such as 'mental retardation', 'neurologic indications'. The genes 'FMR1' and 'SHANK3' are also within the frequently surfac-

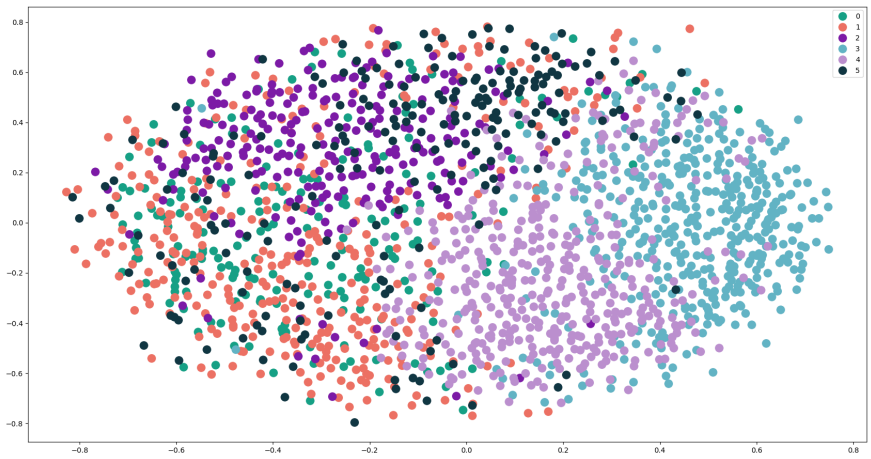

Figure 5.12: K-Means clustering on the full corpus with k=6 showing largely overlapping clusters.

ing terms supporting this, as they are both associated with syndromic co-morbidities [Crawford et al. 2001, Rubeis et al. 2018]. Furthermore, given that the vast majority of these members are within the gene paper collection, we could assume that this cluster is focused on genetic information regarding comorbidities of ASD.

There is some overlap between cluster 0 and 1, and observing the frequently surfacing terms of both clusters, these are various overlapping concepts including 'mental disorders', 'learning disorders', 'neurologic indications' and 'epilepsy'. This can make exactly characterizing the two groupings difficult; however we can observe the mentions of 'attention deficit disorder', 'depressive disorder', 'bipolar disorder', 'anxiety disorders' and 'obsessive compulsive disorder'. Furthermore, concepts which describe symptoms, or are generally related to of these disorders including 'overactivity', 'mood', and 'aggressive behaviour'. We could estimate that this group primarily concerns co-morbid mental disorders within ASDs.

There is a clear difference in the concepts within cluster 2, and clusters 0 and 1. Cluster 2's frequently surfacing concepts are composed of concepts relating to the communication skills phenotype including 'integrated verbal and non-verbal communication', 'ability to converse in social settings', 'reciprocal social interaction', 'ability to convey information' and 'eye contact'. Specific personal traits; 'executive function', 'visual thinking' and 'empathy' are also prominent concepts. Language ability phenotypes appear including 'language skills', 'vocalizations' and 'development or regression of language skills'. We can assume from this that the papers in this group are concerned with the social and behavioural aspects of ASD.

Cluster analysis discovered three major groups of articles which appear to have the common themes of syndromic co-morbid disorders, mental disorders and social-behavioural aspects of ASD. Unfortunately, attempting to view more specific subgroups was inconclusive due to a large overlap. This will likely be due to the restricted number of concepts which they were clustered on. Extending the concept range with more named entities may address this issue, and allow us to identify more specific groupings in the future.

Cluster	Frequently Surfacing Concepts
0 (Green)	Learning disorders, seizures, epilepsy, Fragile-X syndrome, neurologic indications, FMR1, social competency, overactivity, Rett syndrome, SHANK3, schizophrenia, mental disorders, mental retardation, medical history, tuberous sclerosis
1 (Pink)	Attention deficit disorder with hyperactivity, depressive disorder, mental disorder, diagnosis, anxiety disorders, medications, overactivity, learning disorders, cognitive ability, obsessive compulsive disorder, bipolar disorder, mental retardation, Asperger syndrome, mood, aggressive behaviour, epilepsy, neurologic indications
2 (Purple)	Integrated verbal and non-verbal communication, visual thinking, ability to converse in social settings, social competence, reciprocal social interaction, empathy, restricted and repetitive behaviour, Aspergers syndrome, ability to convey information, development or regression of language skills, eye contact, joint attention, vocalisations, language ability, executive function, awareness of social cues, task performance

Table 5.2: Frequently surfacing concepts for each cluster discovered by K-Means.

Chapter 6

Conclusion

6.1 Summary

This project presents a text mining approach to the annotation of ASD-relevant literature, building on previous methods used in information extraction tasks to optimize performance within this domain. We use open source natural language tools, combined with state-of-the-art biomedical entity labelling software, dependency parsing software, a structured ASD ontology utilized as a domain specific vocabulary and manually labelled data to improve performance.

MetaMap's entity labelling tool was improved upon for the context of this task by applying ASD and gene list lookups to achieve a final strict F1-score 0f 0.653 and relaxed F1-score of 0.712, and found it outperformed an existing entity recognition tool by recall and F1-score on a limited set of entities. We achieved strict and relaxed F1-scores of 0.739 and 0.773 respectively, on the identification of ASD-phenotypes from text using nested shallow parsing and stemmed keyword search against a manually constructed dictionary of ASD terms. We developed a method to extract relations between named entities using dependency trees and achieved a strict F1-score of 0.495 and relaxed F1-score of 0.597.

Finally, the resulting system was deployed on a collection of ASD-relevant literature to investigate the concentration of specific ASD-risk genes and phenotypes, identify potential and confirmed gene-phenotype relations, and observe potential subtypes of ASD literature through clustering. We observe a higher representation of syndromic, high confidence or strong candidate genes within the corpus, which can be expected, although suggest this study would likely be more useful for analysing the gene prevalence in a specific type of paper, for example those with language disorder information. We note that comorbidities, in particular mental and learning disorders, are the most prominent phenotype groupings within our corpus, and we note a higher frequency of phenotypes within the 'personal traits' and 'social competence' categories in those considered phenotype papers than papers considered genetic. Executing the relation extraction system on a collection of abstracts identified 183 relations between genes and phenotypes, some of which were suggested while examining the co-occurrence

of gene and phenotype annotations in text. Cluster analysis showed we are able to uncover general groupings of syndromic and mental co-morbid disorders and social-behavioural information within the annotated corpus, however for more specific groupings further improvement is required. In this analysis, we demonstrate some potential uses of such annotation systems to ASD research and present this method as an alternative approach to automated curation in absence of gold-standard data.

6.2 Future Work

There is much room for improvement in each aspect of the system. Evaluation of the system identified issues that occur due to the restrictedness of each module. Perhaps combining the main methodologies proposed here with a statistical approach would address these limitations.

In particular, this project emphasised the need for a gold standard dataset in which to base implementations. Construction of a gold-standard ASD corpus would be highly useful for future ASD research as it would allow access to more sophisticated methods of automated curation.

There is a wide scope for future analysis of the ASD literature. Upon improving annotations to an adequate standard, it would be useful to analyse larger sets of articles to get a more accurate representation of the data. Furthermore, it would be useful to apply the system to clinical data, which may give an insight into prevalence of specific phenotypes actually exhibited by patients with ASD and view trends in those exhibited traits. We show a basic cluster analysis of the annotated literature, however in order to further facilitate access to relevant literature, it would be useful to extend this to perhaps classify papers based on these groupings. It would also be interesting to further analyse the content of papers by applying topic modelling to the literature, such as Latent Dirichlet Allocation, which aims to uncover implicit topics within a corpus using probabilistic modelling on the text [Blei et al. 2003].

Appendix

Phenotype	Highly Co-occurring Genes
Fragile X Syndrome	FMR1, UBE3A, MTOR, NF1
Rett Syndrome	MECP2, MAP2, SCN1A, FOXG1, NSD1, CDKL5, TCF2, MEF2C
Epilepsy	FMR1, PTEN, UB2A, GABRB3, SCN2A, SCN1A, DYRK1A, CHRNA7, NSD1
Schizophrenia	NRXN1, FOXP2, SYNGAP1, CHRNA7, GAD1
Angelman Syndrome	UB3A, MEF2C,
Prader-Willi syndrome	MAGEL2, MAGED1
Tuberous Sclerosis	TSC1, TSC2
Down Syndrome	FAT1, WWOX
Bipolar Disorder	MSN, NRG1, HRAS, SCN8A
Depressive Disorder	NTRK2, HTR2A
Attention Deficit Disorder	ASMT, SCL6A3, PLCB1, SLC9A9
Learning Disorders	SHANK3, MECP2, FMR1, TRIO, PTEN, CNTNAP2, SHANK2
Reciprocal Social Interactions	SHANK3, FMR1, MECP2, PTEN, TRIO, CNTNAP2
Social Competence	SHANK3, SHANK2, SHANK1, OXT, OXTR
Stereotyped, Restricted and Repetitive Behaviour	SHANK3, MECP2, PTEN, CNTNAP2, SHANK2, OXT
Cognitive Ability	TRIO, SHANK3, FMR1, MECP2, PTEN, CNTNAP2
Language Ability	FOXP1, FOXP2, HDAC4, AHl1, TP0, DMPK
Working Memory	BDNF, DC38, HOMER1
Motor Skills	EGR2, CNTN6, NDUFA5
Visual Thinking	APP, CUX1, CNTN5
Perinatal History	DCX, GRIN1, SNAP25, MTHFR, AHL1
Mental Disorders	GABRB3, GRIN2B, RELN, FOXP2
Mental Retardation	NRXN3, DNMT3A, HOXA1
Neurologic Indications	MTOR, SCN1A, BCL2

Table 6.1: ASD Phenotypes and genes with a high co-occurrence.

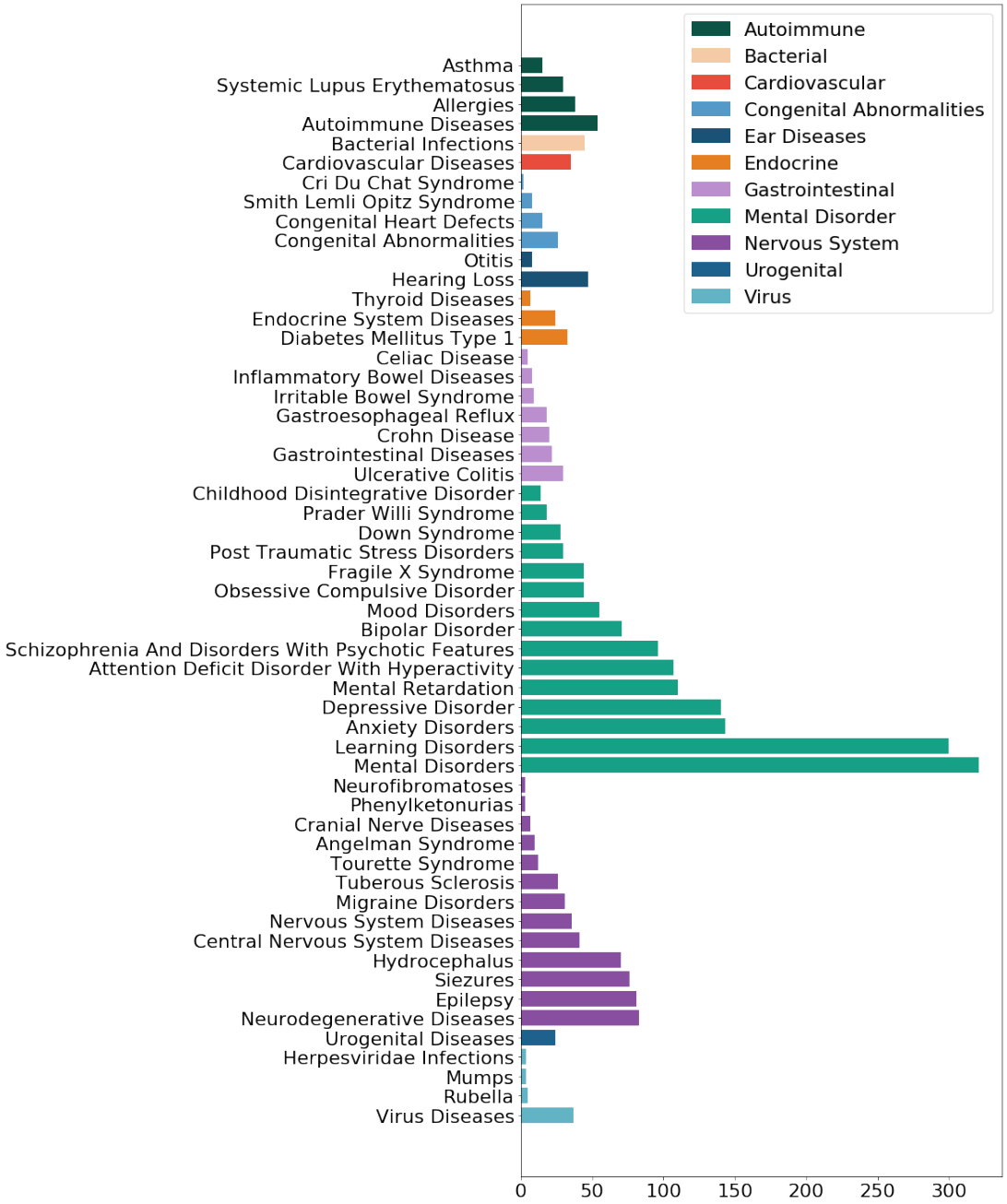

Figure 6.1: TFIDF Counts of comorbidities in the full corpus, grouped by higher level comorbidity types. The comorbidity annotations include general disorder indications such as 'Autoimmune' or 'Mental disorders', as well as specific diseases and disorders.

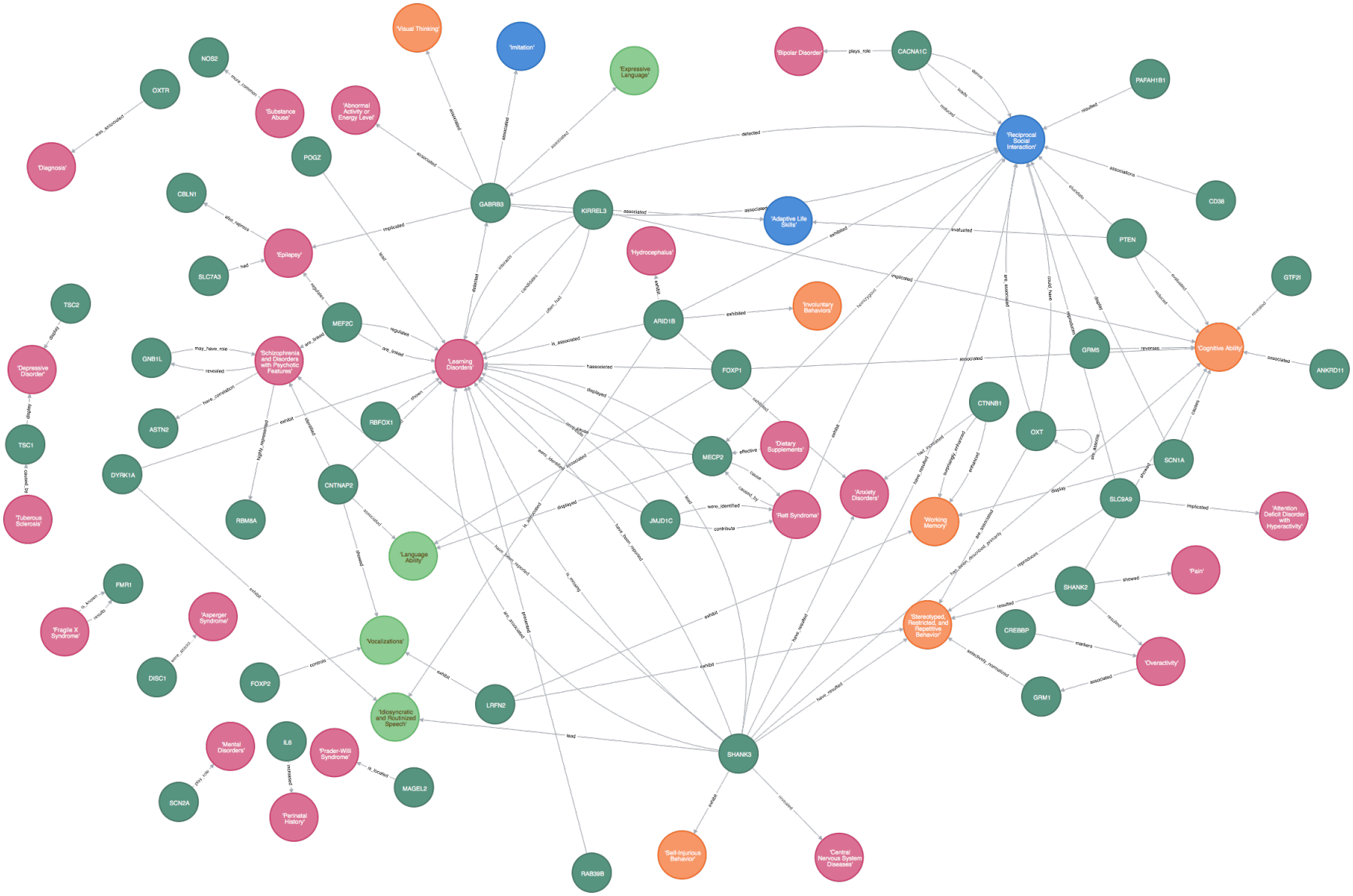

Figure 6.2: Full gene-phenotype relation network extracted from paper abstracts, showing gene instances(dark green), medical history (dark pink), personal trait (light green), and social competence instances (blue), connected by edged labelled with relation found in text.

Bibliography

Al-jawahiri, R. & Milne, E. [2017], Resources available for autism research in the big data era: a systematic review, *in* 'PeerJ'.

Aronson, A. R. & Lang, F.-M. [2010], 'An overview of metamap: historical perspective and recent advances', *Journal of the American Medical Informatics Association* **17**(3), 229236.

ASDPTO - Autism Spectrum Disorder Phenotype Ontology[Online] [Viewed 20 March 2019 [n.d.].
URL: *http://aber-owl.net/ontology/ASDPTO//Browse/*

Bacon, C. & A Rappold, G. [2012], 'The distinct and overlapping phenotypic spectra of foxp1 and foxp2 in cognitive disorders', *Human genetics* **131**, 1687–98.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A. & Hunter, L. E. [2012], 'Concept annotation in the craft corpus', *BMC Bioinformatics* **13**(1), 161.
URL: *https://doi.org/10.1186/1471-2105-13-161*

Baker, S., Ali, I., Silins, I., Pyysalo, S., Guo, Y., Hgberg, J., Stenius, U. & Korhonen, A. [2017], 'Cancer hallmarks analytics tool (chat): a text mining approach to organize and evaluate scientific literature on cancer', *Bioinformatics* **33**(24), 39733981.

Basaldella, M., Furrer, L., Tasso, C. & Rinaldi, F. [2017], 'Entity recognition in the biomedical domain using a hybrid approach', *Journal of Biomedical Semantics* **8**(1).

Batista, D. [n.d.], 'Named-entity evaluation metrics based on entity-level [online] blog [accessed 15 march 2019]'.
URL: *http://www.davidsbatista.net/blog/2018/05/09/Named*$_E$*ntity*$_E$*valuation/*

Björne, j., Kaewphan, S. & Salakoski, T. [2013], Uturku: Drug named entity recognition and drug-drug interaction extraction using svm classification and domain knowledge.

Blei, D. M., Ng, A. Y. & Jordan, M. I. [2003], 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**, 993–1022.
URL: *http://dl.acm.org/citation.cfm?id=944919.944937*

Bryson, S. [1997], 'Epidemiology of autism: Overview and issues outstanding', *Handbook of autism and pervasive developmental disorders* pp. 41–46.

Caporaso, J. G., Baumgartner, W. A., Randolph, D. A., Cohen, K. B. & Hunter, L. [2007], 'Mutationfinder: a high-performance system for extracting point mutation mentions from text', *Bioinformatics* **23**(14), 18621865.

Centers for Disease Control and Prevention [2018], 'Autism and developmental disabilities monitoring network: Community report on autism'.
URL: *https://www.cdc.gov/ncbddd/autism/addm-community-report/documents/addm-community-report-2018-h.pdf*

Chinchor, N. & Sundheim, B. [1993], Muc-5 evaluation metrics, *in* 'Proceedings of the 5th Conference on Message Understanding', MUC5 '93, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 69–78.
URL: *https://doi.org/10.3115/1072017.1072026*

Crawford, D., Acua, J. & Sherman, S. [2001], 'Fmr1 and the fragile x syndrome: Human genome epidemiology review', *Genetics in medicine : official journal of the American College of Medical Genetics* **3**, 359–71.

Croen, L. A., Zerbo, O., Qian, Y., Massolo, M. L., Rich, S., Sidney, S. & Kripke, C. [2015], 'The health status of adults on the autism spectrum', *Autism* **19**(7), 814823.

Cunha, C., Brambilla, R. & Thomas, K. [2010], 'A simple role for bdnf in learning and memory?', *Frontiers in molecular neuroscience* **3**, 1.

D. Karp, P. [2016], 'Can we replace curation with information extraction software?', *Database: The Journal of Biological Databases and Curation* **2016**.

de Marnee, M.-C. & Manning, C. [2008], 'Stanford typed dependencies manual'.

de Sena Cortabitarte, A., Degenhardt, F., Strohmaier, J., Lang, M., Weiss, B., Roeth, R., Giegling, I., Heilmann-Heimbach, S., Hofmann, A., Rujescu, D. & et al. [2017], 'Investigation of shank3 in schizophrenia'.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28371232*

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators; Centers for Disease Control and Prevention [2010], 'Prevalence of autism spectrum disorder among children aged 8 years', *MMWR Surveill Summ.* pp. 63:1–21.

Devlin, B. & Scherer, S. [2012], 'Genetic architecture in autism spectrum disorder', *Current opinion in genetics development* **22**, 229–37.

Döhling, L. & Leser, U. [2011], 'Equatornlp: Pattern-based information extraction for disaster response', *CEUR Workshop Proceedings* **798**.

Elayavilli, R. K., Rastegar-Mojarad, M. & Liu, H. [2017], Belminer information extraction system to extract bel relationships.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcn, C., Montiel-Nava, C., Patel, V., Paula, C. S., Wang, C. & et al. [2012], 'Global prevalence of autism and other pervasive developmental disorders'.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763210/*

Elsevier [2018], 'Automated vs manual literature curation: Extracting more information from scientic literature'.
URL: *https://www.elsevier.com/*$_{data/assets/pdf_file/0007/97036/Automated-vs-Manual-WEB_new.pdf}$

Fluck, J. & Hofmann-Apitius, M. [2013], 'Text mining for systems biology', *Drug discovery today* **19**.

Foundation, N. S. [2018], 'Science engineering indicators 2018'.
URL: *https://nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/highlightsoutputs-of-s-e-research-publications*

Foundation, S. [2019], 'About the gene scoring module'.
URL: *https://gene.sfari.org/about-gene-scoring/*

Fundel-Clemens, K., Kffner, R. & Zimmer, R. [2007], 'Relex - relation extraction using dependency parse trees', *Bioinformatics (Oxford, England)* **23**, 365–71.

Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. [2015], 'The role of ontologies in biological and biomedical research: a functional perspective', *Briefings in Bioinformatics* **16**(6), 1069–1080.
URL: *https://dx.doi.org/10.1093/bib/bbv011*

Graff, H., Berkeley, S., Evmenova, A. & L. Park, K. [2014], 'Trends in autism research: A systematic journal analysis', *Exceptionality: A Special Education Journal* **22**, 158–172.

Hara, M., Ohba, C., Yamashita, Y., Saitsu, H., Matsumoto, N. & Matsuishi, T. [2015], 'De novo shank3 mutation causes rett syndrome-like phenotype in a female patient', *American journal of medical genetics. Part A* **167**.

Harmston, N., Filsell, W. & Stumpf, M. P. [2010], 'What the papers say: Text mining for genomics and systems biology', *Human Genomics* **5**(1), 17.

Hewitson, L. [2013], 'Scientific challenges in developing biological markers for autism', *OA Autism* **1**(1).

Human Genome Variation Society [2016], 'Sequence variant nomenclature'. Accessed: 2019.
URL: *http://varnomen.hgvs.org/*

J. Tanaka, A., Cho, M., Willaert, R., Retterer, K., Zarate, Y., Bosanko, K., Stefans, V., Oishi, K., Williamson, A., N. Wilson, G., Basinger, A., Barbaro-Dieber, T., Ortega, L., Sorrentino, S., K. Gabriel, M., J. Anderson, I., Ferrin, M. A., E. Schnur, R. & Chung, W. [2017], 'De novo variants in ebf3 are associated with hypotonia, developmental delay, intellectual disability, and autism', *Molecular Case Studies* **3**, a002097.

Jayanthi, S. K. & Kavi Priya, C. [2018], 'Clustering approach for classification of research articles based on keyword search', *International Journal of Advanced Research in Computer Engineering and Technology* **7**(1).
URL: *http://ijarcet.org/wp-content/uploads/IJARCET-VOL-7-ISSUE-1-86-90.pdf*

Jimeno-Yepes, A., Jiménez-Ruiz, E., Lee, V., Gaudan, S., Llavori, R. B. & Rebholz-Schuhmann, D. [2008], 'Assessment of disease named entity recognition on a corpus of annotated sentences', *BMC Bioinformatics* **9**, S3 – S3.

Katona, M. & Farkas, R. [2014], Szte-nlp: Clinical text analysis with named entity recognition, *in* 'Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)', Association for Computational Linguistics, Dublin, Ireland, pp. 615–618.
URL: *http://www.aclweb.org/anthology/S14-2108*

Khordad, M. & Mercer, R. E. [2017], 'Identifying genotype-phenotype relationships in biomedical text', *Journal of Biomedical Semantics* **8**(1), 57.
URL: *https://doi.org/10.1186/s13326-017-0163-8*

Kim, S., Yeganova, L. & Wilbur, W. [2016], 'Meshable: Searching pubmed abstracts by utilizing mesh and mesh-derived topical terms', *Bioinformatics* **32**, btw331.

Lai, C. S. L., Fisher, S., Hurst, J., Vargha-Khadem, F. & Monaco, A. [2001], 'Lai, c. s. l., fisher, s. e., hurst, j. a., vargha-khadem, f. monaco, a. p. a forkhead-domain gene is mutated in a severe speech and language disorder. nature 413, 519-523', *Nature* **413**, 519–523.

Lévy, F., Tomeh, N. & Ma, Y. [2014], 'Ontology-based technical text annotation'.

Li, L., Hu, L., Xiong, S., Xing, W., Yuan, X., Fu, Y., Peng, J., Qi, J. & Zhang, X. [2018], 'A genephenotype relationship extraction pipeline from the biomedical literature using a representation learning approach', *Bioinformatics* **34**(13), i386–i394.
URL: *https://doi.org/10.1093/bioinformatics/bty263*

Lim, S. & Kang, J. [2018], 'Chemicalgene relation extraction using recursive neural network', *Database* **2018**.
URL: *https://doi.org/10.1093/database/bay060*

Lipton, J. O. & Sahin, M. [2014], 'The neurology of mtor', *Neuron* **84**(2), 275 – 291.
URL: *http://www.sciencedirect.com/science/article/pii/S0896627314008927*

Lobo, M., Lamurias, A. & Couto, F. [2017], 'Identifying human phenotype terms by combining machine learning and validation rules', *BioMed Research International* **2017**, 1–8.

Luksic, M. M., Urbancic, T., Petric, I. & Cestnik, B. [2016], 'Autism research dynamic through ontology-based text mining', *Advances in Autism* **2**(3), 131–139.
URL: *https://doi.org/10.1108/AIA-01-2016-0001*

M Cohen, A. & Hersh, W. [2005], 'A survey of current work in biomedical text mining', *Briefings in bioinformatics* **6**, 57–71.

Majumdar, J., Naraseeyappa, S. & Ankalaki, S. [2017], 'Analysis of agriculture data using data mining techniques: application of big data', *Journal of Big Data* **4**(1).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. [2014], The Stanford CoreNLP natural language processing toolkit, *in* 'Association

for Computational Linguistics (ACL) System Demonstrations', pp. 55–60.
URL: *http://www.aclweb.org/anthology/P/P14/P14-5010*

McCray, A. T., Trevvett, P. & Frost, H. R. [2013], 'Modeling the autism spectrum disorder phenotype', *Neuroinformatics* **12**, 291–305.

Mcpartland, J. & Volkmar, F. R. [2012], 'Autism and related disorders', *Neurobiology of Psychiatric Disorders Handbook of Clinical Neurology* p. 407418.

Munkhdalai, T., Li, M., Batsuren, K., Ah Park, H., Hyeon Choi, N. & Ryu, K. [2015], 'Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations', *Journal of cheminformatics* **7**, S9.

Rasmussen, J., Graham, A. M., Entringer, S., Gilmore, J. H. & Buss, C. [2019], 'Maternal interleukin-6 concentration during pregnancy is associated with variation in frontolimbic white matter and cognitive development in early life', *NeuroImage* **185**, 825–835.

Ravi, S. & Sundarambal, M. [2018], 'Clustering of biomedical documents using ontology-based tf-igm enriched semantic smoothing model for telemedicine applications', *Cluster Computing* .

Rubeis, S. D., Siper, P. M., Durkin, A., Weissman, J., Muratet, F., Halpern, D., Trelles, M. D. P., Frank, Y., Lozano, R., Wang, A. T. & et al. [2018], 'Delineation of the genetic and clinical spectrum of phelan-mcdermid syndrome caused by shank3 point mutations', *Molecular Autism* **9**(1).

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H. & Reichenberg, A. [2017], 'The Heritability of Autism Spectrum Disorder: Reassessing the Heritability of Autism Spectrum Disorders Letters', *JAMA* **318**(12), 1182–1184.
URL: *https://dx.doi.org/10.1001/jama.2017.12141*

Sasaki, Y., Tsuruoka, Y., McNaught, J. & Ananiadou, S. [2008], 'How to make the most of ne dictionaries in statistical ner', *BMC Bioinformatics* **9**(11), S5.
URL: *https://doi.org/10.1186/1471-2105-9-S11-S5*

Schäfer, U. [2006], Ontonerdie – mapping and linking ontologies to named entity recognition and information extraction resources, pp. 1756–1761.

Singhal, A., Leaman, R., Catlett, N., Lemberger, T., McEntyre, J., Polson, S., Xenarios, I., Arighi, C. & Lu, Z. [2016], 'Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges', *Database* **2016**(0).

Soomro, P. D., Kumar, S., Shaikh, A. A. & Raj, H. [2017], Bio-ner : Biomedical named entity recognition using rule-based and statistical learners.

The UMLS Semantic Network [2009]. Accessed: 2016-07-26.
URL: *https://semanticnetwork.nlm.nih.gov/*

Tick, B., Bolton, P., Happe, F., Rutter, M. & Rijsdijk, F. [2015], 'Heritability of autism spectrum disorders: A meta-analysis of twin studies', *Journal of Child Psychology and Psychiatry* **57**, n/a–n/a.

Wang, J., Lin, Z.-J., Liu, L., Xu, H.-Q., Shi, Y.-W., Yi, Y.-H., He, N. & Liao, W.-P. [2017], 'Epilepsy-associated genes', *Seizure* **44**, 11 – 20. 25th Anniversary Issue.
URL: *http://www.sciencedirect.com/science/article/pii/S1059131116302989*

Wang, M., Xia, H., Sun, D., Chen, Z., Wang, M. & Li, A. [2014], 'Literature mining of protein phosphorylation using dependency parse trees', *Methods* **67**(3), 386393.

Wei, C.-H., Harris, B., Li, D., Z Berardini, T., Huala, E., Kao, H.-Y. & lu, Z. [2012], 'Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in pubmed abstracts', *Database : the journal of biological databases and curation* **2012**, bas041.

Wei, C.-H., Kao, H.-Y. & Lu, Z. [2013], 'Pubtator: a web-based text mining tool for assisting biocuration', *Nucleic Acids Research* **41**.
URL: *http://nar.oxfordjournals.org/content/early/2013/05/22/nar.gkt441.abstract?keytype=refijkey=mj3Ee*

Xu, H., Hu, C. & Shen, G. [2009], 'Discovery of dependency tree patterns for relation extraction', **2**.

Yates, B., Braschi, B., Gray, K., Seal, R., Tweedie, S. & Bruford, E. [2017], 'Genenames.org: the hgnc and vgnc resources in 2017'. Accessed: 2019.

Yiu, E. M. & Baets, J. [2015], Chapter 16 - congenital and early infantile neuropathies, *in* B. T. Darras, H. R. Jones, M. M. Ryan & D. C. D. Vivo, eds, 'Neuromuscular Disorders of Infancy, Childhood, and Adolescence (Second Edition)', second edition edn, Academic Press, San Diego, pp. 289 – 318.
URL: *http://www.sciencedirect.com/science/article/pii/B9780124170445000160*

Yong An, J. & Claudianos, C. [2016], 'Genetic heterogeneity in autism: From single gene to a pathway perspective', *Neuroscience and biobehavioral reviews* **68**.

Yu, K., Lung, P.-Y., Zhao, T., Zhao, P., Tseng, Y.-Y. & Zhang, J. [2018], 'Automatic extraction of protein-protein interactions using grammatical relationship graph', *BMC Medical Informatics and Decision Making* **18**.

Zhang, S., Kang, T., Qiu, L., Zhang, W., Yu, Y. & Elhadad, N. [2016], Cataloguing treatments discussed and used in online autism communities, Vol. 2017.

Zhou, D., Zhong, D. & He, Y. [2014], 'Biomedical relation extraction: From binary to complex', *Computational and mathematical methods in medicine* **2014**, 298473.

Zhu, W., Li, J., Chen, S., Zhang, J., Vetrini, F., Braxton, A., Eng, C. M., Yang, Y., Xia, F., Keller, K. L. & et al. [2018], 'Two de novo novel mutations in one shank3 allele in a patient with autism and moderate intellectual disability', *American Journal of Medical Genetics Part A* **176**(4), 973979.

Zylstra, R., Prater, C. D., Walthour, A. E. & Feliciano Aponte, A. [2014], 'Autism: Why the rise in rates? our improved understanding of the disorder and increasingly sensitive diagnostic tools are playing a role–but so are some other factors.', *Journal of Framily Practice* pp. 316–3.