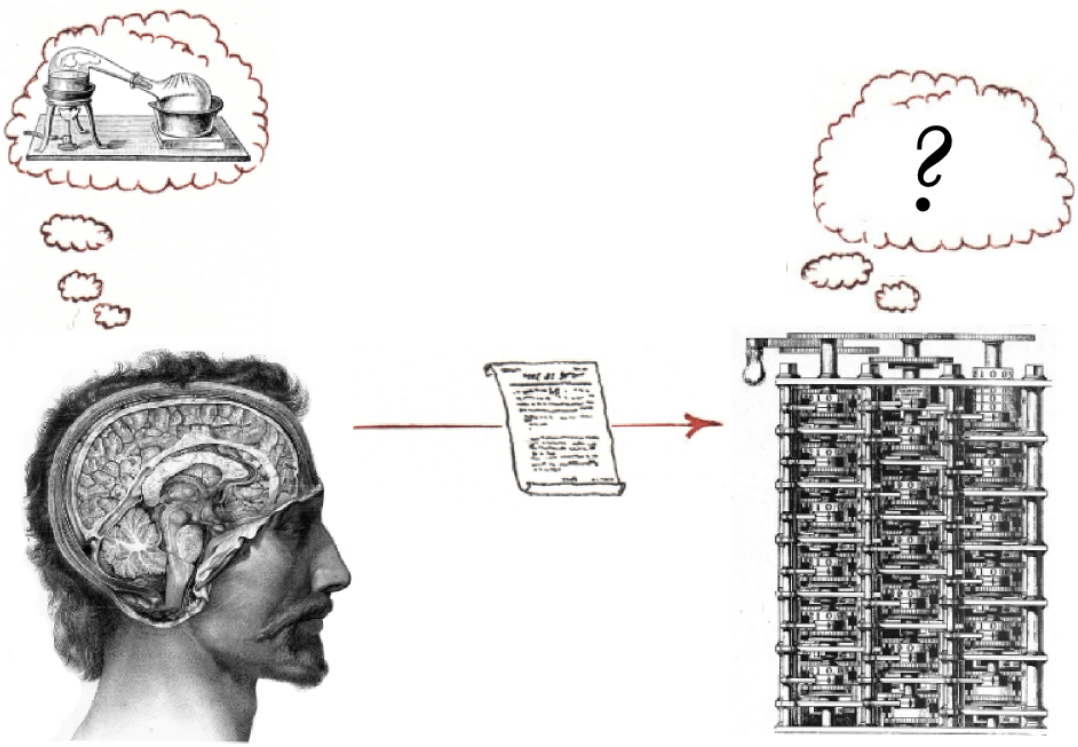
USING NATURAL LANGUAGE PROCESSING METHODS TO SUPPORT CURATION OF A CHEMICAL ONTOLOGY.

ADAM BERNARD



Homerton College, University of Cambridge
&
European Bioinformatics Institute

May 2014

This dissertation is submitted for the degree of Doctor of Philosophy.

Adam Bernard: *Using Natural Language Processing methods to support curation of a chemical ontology.,* Doctor of Philosophy, May 2014

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the word limit as specified by the Degree Committee for the Faculty of Biology.

Cambridge, May 2014

Adam Bernard

To Hugh R. S. Jones, who taught me the value of back-of-the-envelope calculations, and to the memory of my grandparents Elise Kersh and Sidney & Norma Bernard.

SUMMARY

Adam Bernard

Using Natural Language Processing methods to support curation of a chemical ontology.

This thesis describes various techniques to assist the curation of a chemical ontology (ChEBI) using a combination of textmining techniques and the resources of the ontology itself. ChEBI is an ontology of small molecules that are either produced by, or otherwise relevant to, biological organisms. It is manually expert-curated, and as such has high reliablity but incomplete coverage. To make efficient use of curator time, it is desirable to have automatic suggestions for chemical species and their properties, to be assessed for inclusion in ChEBI.

Having developed a system to identify chemicals within biological text, I use a combination of a syntactic parser and a small set of textual patterns to extract hypernyms of these chemicals (categories of chemicals where there is an is-a relationship e.g. glycine is-a amino acid) where both the chemical and the hypernym can be resolved to entities already within ChEBI. I identify features that affect the confidence we can have in the assignment of these hypernyms, and use these to develop a classifier so that the more certain hypernyms can be filtered.

The system to identify hypernyms is extended to identify non-hypernymic relationships; the patterns used to extract these are informed by some of the shortcomings of the hypernym resolution. These relationships connect chemicals not only to other chemicals but also to concepts — such as diseases, proteins, and cellular components — from other biological ontologies. Different relation types connect chemicals to different types of concept, and this can be used to improve detection of incorrectly-extracted relations.

I characterize these properties and demonstrate that it is possible, using the chemicals that they describe as features, to infer relations between properties. I assess the reliability of these inferred relations.

ACKNOWLEDGMENTS

The annotation in chapters 4 and 6 was performed by Gareth Owen and Steve Turner from the ChEBI team, who together with Janna Hastings and Paula de Matos also provided help with understanding the workings of the ChEBI project.

The annotation in chapter 7 was performed by Peter Corbett and Colin Batchelor, to whom many thanks for giving up their time at short notice.

Technical advice was given by Peter Corbett (who *inter alia* provided a training corpus for the named-entity recognition system), Ian Lewin, Simone Teufel, and Hanna Wallach.

Clare Boothby provided invaluable organizational help and encouragement, as well as proof-reading.

Colin Batchelor was a huge help in providing both technical advice and logistical and moral support, all of which made a tremendous difference and without which I would not have completed this project.

My Thesis Advisory Committee consisted of my supervisor Dietrich Rebholz-Schuhmann, along with Henning Hermjakob, Reinhardt Schneider, and Simone Teufel. My tutor at Homerton College was Penny Barton.

Simone Teufel read various drafts of this thesis, and was immensely helpful in giving detailed feedback and advice, especially in helping me organize the structure of the thesis and process the annotation results.

The project was funded by the BBSRC with Pfizer UK.

The Systems group at the EBI kept the hardware and and software for the project running smoothly, and Ian Jackson administered the server used for annotations and backup.

During medical mishaps in the course of the project, I was patched up repeatedly by the staff at Addenbrooke's Hospital and supported by Jennifer Koenig and the University of Cambridge's Disability Resource Centre, and the staff at Huntingdon Road GP surgery, especially Dr Karen Newman.

Many friends and colleagues provided enormous amounts of support and encouragement: thanks especially (in addition to those above) to Kathryn Taylor, Tom Womack, Bridget Bradshaw, Anika Oellrich, Peter Corbett, Heather Hooper, Rachel Coleman, Jack Vickeridge, Emily Divinagracia, and my parents Robert & Gill Bernard.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BNC British National Corpus

ChEBI Chemical Entities of Biological Interest

DDC Dewey Decimal Classification

DO Human Disease Ontology

EBI European Bioinformatics Institute

GO Gene Ontology

GOA Gene Ontology Annotation

HMM Hidden Markov Model

IAA Interannotator Agreement

KD Knowledge Discovery

LCC Library of Congress Classification

LSA Latent Semantic Analysis

LSP Lexicosyntactic Pattern

MeSH Medical Subject Headings

ML Machine Learning

NER Named Entity Recognition

NLP Natural Language Processing

NP Noun Phrase

NP2 Two-part Noun Phrase

OBO Open Biological Ontologies

PMI Pointwise Mutual Information

PMID PubMed ID

POS Part of speech

PRO Protein Ontology

SQL Structured Query Language

SVM Support Vector Machine

UMLS Unified Medical Language System

1

INTRODUCTION

Given a set of knowledge about the world, or about a particular part of the world, how are we to organize it? One method is simply to lay facts out on paper (or, more recently, on its electronic equivalent) as essays, and delegate the problem of building a coherent understanding of the world to the reader, perhaps relying on their innate understanding of the world and their learning to provide scaffolding on which they can put the facts into context.

A slightly more accessible approach would be to produce gazetteers of particular *types* of fact. For examples of such collations, we can see dictionaries, almanacs, atlases, bestiaries, and herbals dating back at least to Pliny the Elder's *Naturalis Historiae*[*].

A starting point for a more *structured* approach came in 1668 with John Wilkins' brilliant yet quixotic *An essay towards a real character and a philosophical language*[112], which lays out in its introduction the problems with human languages as a means to describe the world, and adopts the principle of a hierarchical organization of concepts — relations, actions, properties, and objects. Notably, the same scheme contains both the abstract and the concrete members of the hierarchy — such as *Castle*, *Sneezing*, *Right Angle*, *Future*, and *Length* — and the various terms that define the types of entity and the relations between them — such as *Equality*, *Quantity*, *Action*. A small section of the scheme can be seen in Figure 1. Wilkins' aim for the project was to suggest a new and more logically founded language for humans to speak; in this respect he met with complete lack of interest, and the work was doomed to relative obscurity.

In 1735, Carl Linnaeus published the first edition of the influential *Systema Naturae*[62]. This work had a narrower scope than Wilkins', covering the natural world, divided into the animal, vegetable, and mineral "kingdoms". It set the direction of classification in the natural sciences, and its mode of working was rapidly adopted. It concerned itself with hierarchical ordering of species, but did not contain other relationships between them.

In modern terms, we would describe Linnaeus' scheme as a *taxonomy*, and Wilkins' as an *ontology*. The former is fully describable as a tree, where

* http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.02.0138:toc

each entity's position is described by its *parent*. We shall call this type of relationship `is_a` *relationships* or *hypernymy*[64] — so we can say that *Lynx* `is_a` *Cat* or *"Cat" is a hypernym of "Lynx"** . The latter work has more varied relationships, so while *Wolf* `is_a` *Dog-Kind*, it is also specified that *Wolf* `is` *Terrestrial* (which is glossed as related to Land)† and *Wolf* `has_property` *Wildness*. It follows that a taxonomy can be seen as a special case of an ontology, which just happens only to have the one kind of relation — see for example Figure 2, and in which it just happens that each entry has precisely one parent. Figure 3 has three relation types - `is_a`, `has_part`, and `inhabits`. When talking about these structures, we will refer to the points that represent the entities as *nodes* and the lines connecting them as *edges*.

There has been much recent interest in the field of biomedical ontologies; from the Gene Ontology (GO)[11] around the turn of the last decade, the field has seen a great deal of development[98], with ontologies becoming an accepted tool for arranging and describing the rapidly-growing body of biomedical knowledge.

* The inverse of this relationship is hyponymy: *"Lynx" is a hyponym of "Cat"*

† Wilkins divides Dog-kinds into the Terrestrial, which along with the wolf are the dog, the fox, and the badger, and the Aquatic, in which category he places the seal and the morse (walrus)

64 *Stone.* Part. II.

V. EARTHY CONCRETIONS DISSOLVIBLE.

V. Such EARTHY CONCRETIONS as commonly grow *in Mines*, together with ſuch *other factitious Subſtances* as have ſome analogy to theſe, and are DISSOLVIBLE by Fire or Water, may be diſtinguiſhed by their being

- *Not inflammable:*
  - *More ſimple*; being ſeveral kinds of *Salt*, ‖ whether of the *Sea-water*, the moſt neceſſary *Condiment for Meat*; or of the *Air*, *uſed* as a chief ingredient *in* the *making of Gunpowder*.
    1. SALT, *Brine.* / NITRE, *Salt-peter.*
  - *Earth*; ‖ of a *ſtyptic* quality and abſterſive, proper for the drying of Wounds, commonly *boiled up* into a conſiſtence *from a mineral water*; or that other kind of Earthy Salt *dug up in* great *lumps*.
    2. ALUME. / SAL GEMMÆ.
  - *Metals* of all kinds, ſometimes called Sugars and Cryſtals; but agreeing in the common nature with that which is ſtyled
    3. VITRIOL, *Chalchanthus*, *Copperas.*
  - *Vegetables*; made ‖ either *by fermentation*, or *by burning*.
    4. TARTAR. / ALCALI.
  - *Animal Subſtances*, made *by Diſtillation*, called
    5. URINOUS SALT.
  - *More mixed* of other Salts; ‖ more *volatile*, or *fixed*.
    6. SAL AMMONIAC. / CHRYSOCOLLA, *Borax.*
- *Inflammable*; of a more
  - *Dry* conſiſtence, and *Yellowiſh* colour.
    7. SULPHUR, *Brimſtone.*
  - *Clammy* and tenacious conſiſtence
    - *Not ſweet*-ſented; ‖ more *ſolid*, or more *liquid*.
      8. BITUMEN-*inous.* / NAPHTHA.
    - *Sweet*-ſented.
      9. AMBERGRIS.

Figure 1: Part of John Wilkins' 1668 general-domain ontology covering chemicals. The hierarchical structure is clearly visible, as is the use of a controlled vocabulary, the canonical name being in upper-case with synonyms and derivatives following each entry in italics.[112]

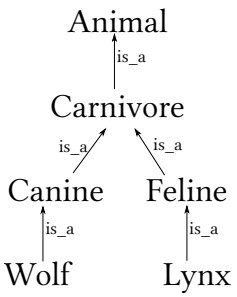


Figure 2: A tiny taxonomy, loosely based on Linnaeus[62]

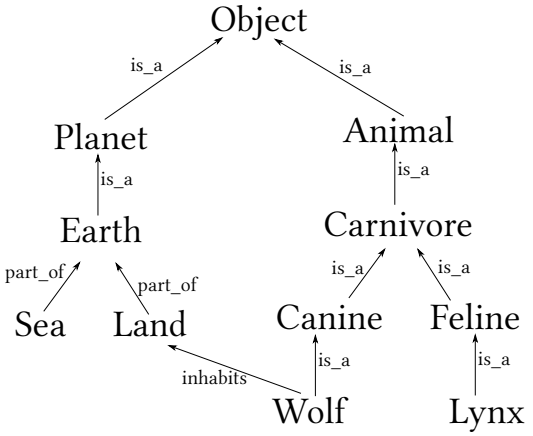


Figure 3: A tiny ontology, loosely based on Wilkins[112]

1.1 USES OF ONTOLOGIES

Ontologies have a variety of uses, and new uses are continually being found.

At the most basic level, they can be used for manual reference by humans, effectively as an encyclopaedia. They may also provide a framework for organizing an encyclopaedia or other reference work — Roget's *Thesaurus* (1805) is an early example of such a work[55] — or physical resources such as libraries; compare taxonomies such as the Dewey Decimal Classification (DDC) (1876) and its younger sister the Library of Congress Classification (LCC) (1897), which are some of the earlier and best-known instances of taxonomies.

A more technical use of ontologies is their use as a controlled vocabulary. See for example Rogers(1963)[91], which details the rationale behind the collation of Medical Subject Headings (MeSH)* to facilitate ease of indexing. Controlled vocabularies can avoid the ambiguity of, for instance, the word *nucleus* having the senses both of *cell nucleus* and *atomic nucleus*.

Another application of ontologies is for inference. One of the major uses of GO annotations is in characterizing sets of genes, such as may arise from

* http://www.ncbi.nlm.nih.gov/mesh

the results of a microarray experiment. In these experiments, the expression of some genes may be found to be upregulated in response to a stimulus such as a disease, a drug, or an environmental perturbation. It is often desirable to know what properties are shared by these genes as compared to genes in general. There are many tools that identify such enrichment or over-representation, a selection of which are listed at the GO website.* The taxonomic nature of GO is vital for this process since it allows conclusions to be drawn about properties that are not necessarily the leaf nodes with which genes have been annotated.

Ontologies can also be used for inference of new biological knowledge, as for example Bodenreider et al.(2005)[18] have done within the domain of GO.

1.2 AUTOMATIC POPULATION OF ONTOLOGIES

It should go without saying† that there is a lot of information out there.‡ More specifically for our purposes, there is a huge amount of biological information out there. As of August 2012, PubMed contained records for over 22 million articles§ with just over 2.5 million of these available as full-text articles in PubMed Central¶.

Relatedly, there is an increasing number of computer-aware biologists, a blooming industry in informatics, bioinformatics, chemoinformatics, and so on, and ever-faster computers that can make use of more and more information. The demand for structured information, as well as the supply of unstructured information, is growing apace.

Whilst the ontology-compilers before the advent of the modern computer had little choice but to employ humans to create their resources, in the last few decades there has been an interest in the extent to which computers might be employed instead. Whereas John Wilkins created his ontology with the aim of establishing the foundation of a new language, today we look to processing natural languages (still with all the infelicities for this purpose that Wilkins noted, of ambiguity, synonymy, and lack of clarity) to provide a

* http://www.geneontology.org/GO.tools.microarray.shtml

† Although this thesis, in common with the prefaces of every other work in the field of bioinformatics, will say it anyway.

‡ http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm provides a somewhat dated (2002) attempt at quantifying information generated by humans. They estimate (*inter alia*) that the text in all US academic research libraries came to about 2 Petabytes (2×10^{15} bytes), equivalent to 10^{12} typewritten pages; this in turn is dwarfed by the volume of information estimated to be generated annually in electronic format. They further estimate that stored information was increasing at 30% per year.

§ http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&cmd=search&term=1800:2100[dp]

¶ http://www.ncbi.nlm.nih.gov/pmc/?term=1800:2100[dp]

basis for our ontologies. For some purposes the results are "good enough", for others the barriers of the irregularities of natural language provide an obstacle to the technology such that human curation is still the best solution.

We must also be aware that as well as the different purposes to which the ontology may be put, there are also other differences between domains that affect the results that text-processing can achieve. One factor is the availability of information, and the simplicity with which statements are made. Another is the degree to which the jargon is specialised, such that tools developed for general-purpose text may need adapting to work well with domain-specific text. A third factor is the complexity of the structure of the information, and yet another is the degree of unanimity or objectiveness with which the information is stated.

To illustrate these points, consider an example at the more straightforward end of the spectrum, and suppose we have a list of nation-states and wish to identify their capital cities. We can define a model where there is just one type of relation: `has_capital`, each state has one capital, and, with few exceptions, there is no disagreement on what the capital is. We might wish to limit ourselves to recent documents, since we know that nations can re-designate their capitals from time to time. We can probably use text that is written in a reasonably simple style, and if we are lucky, we might find a single document (such as an encyclopaedia) or collection of documents that will be written very consistently — perhaps even taking care to state the fact in the same way every time. We know that all the capitals will necessarily be cities, and we know when we have found the complete set. If we have more than one candidate for the role of capital for a country, we can rank the contenders and choose the best one.

Biological and chemical data, in contrast, have few of these advantages. One positive point is that there is at least a reasonable collection of literature in electronic format and written in reasonably formal English. Weighed against this, there is much more subjectivity and less unanimity — both about the types of relations that it is worthwhile and meaningful to define, and about whether any given example of that relation holds true. The language used is complex — compare, for instance "*Paris*" and "*Vascular endothelial growth factor receptor antagonist*"*. Furthermore, there tend to be many synonyms for the same concepts — and sometimes multiple concepts that can share the same name or (especially) abbreviation. The way that a

* Some questions that are hard for even a human fluent in English to answer: is it the growth factor, the receptor, or the antagonist which is vascular? Is "*endothelial growth factor receptor*" anything that is a receptor for any growth factor produced in the endothelium (or any receptor produced in the endothelium for a growth factor), or a name given to a specific protein?

biological molecule behaves in one circumstance may be very different to its behaviour under another circumstances — for an extreme example, a comment that water is lethal in small volumes might be perfectly true and meaningful in a discourse where the topic is drowning, but how to codify this fact in a useful way is not at all obvious.

1.3 CHEBI

Chemical Entities of Biological Interest (ChEBI)[40] is a reliable expert-curated ontology that contains chemicals, properties of chemicals, and uses of chemicals, with a focus on those molecules that are important within the field of biology. It covers "small" molecules (*i.e.* not routinely including macromolecules such as nucleic acids and proteins) as well as atoms and subatomic particles.

It is used both as a human-readable reference work and as a machine-readable data source. In this latter category it provides data suitable for automatic inference across the domain of chemicals in biology[17]. It has coverage of chemicals of biological interest in general, as opposed to being, for instance, drug-specific (such as DrugBank or the Anatomical Therapeutic Chemical list). The ontology consists of four sub-ontologies: Molecular Structure, Biological Role, Application, and Subatomic Particle.

To help ensure reliability, ChEBI is manually curated by experts. Each assertion within the ChEBI ontology is required to be supported by evidence from the scientific literature. Consequently, while it tends to be very accurate, the ontology is laborious to extend, and hence it is far from complete — there is much appropriate information in the published literature that has yet to be integrated into the ontology.

It is rich in structure as opposed to a pure taxonomy; the example in Listing 1 has relations of type `is_a`, `has_role`, and `has_part`. In this case it `is_a` CHEBI:36807 (hydrochloride), `has_part` CHEBI:3699 (cimetidine), and `has_role` CHEBI:49201 (anti-ulcer drug) — shown in Listing 2, the last line of which indicates that CHEBI:49201 (anti-ulcer drug) `is_a` CHEBI:23888 (drug).

Adding edges is time-consuming and there is no easy way of telling when edges are missing — a chemical entity may, for instance, have any number of roles.

```
[Term]
id: CHEBI:50362
name: cimetidine hydrochloride
def: "A hydrochloride that has formula C10H17ClN6S." []
synonym: "Tagamet" RELATED BRAND_NAME [DrugBank:]
synonym: "Cimetidine HCl" RELATED [ChemIDplus:]
synonym: "C10H17ClN6S" RELATED FORMULA [ChEBI:]
synonym: "C10H16N6S.HCl" RELATED FORMULA [KEGG DRUG:]
synonym: "[H+].[Cl-].CN\\C(NCCSCc1nc[nH]c1C)=N\\C#N" RELATED SMILES [ChEBI:]
synonym: "InChI=1S/C10H16N6S.ClH/c1-8-9(16-7-15-8)5-17-4-3-13-10(12-2)14-6-11;/h7H
     ,3-5H2,1-2H3,(H,15,16)(H2,12,13,14);1H" RELATED InChI [ChEBI:]
synonym: "InChIKey=QJHCNBWLPSXHBL-UHFFFAOYSA-N" RELATED InChIKey [ChEBI:]
xref: ChemIDplus:70059-30-2 "CAS Registry Number"
xref: DrugBank:DB00501 "DrugBank"
xref: KEGG DRUG:D03503 "KEGG DRUG"
relationship: has_part CHEBI:3699
is_a: CHEBI:36807
relationship: has_role CHEBI:49201
```

Listing 1: An entry for a chemical — cimetidine hydrochloride — within ChEBI

```
[Term]
id: CHEBI:49201
name: anti-ulcer drug
def: "One of various classes of drugs with different action mechanisms used to
     treat or ameliorate peptic ulcer or irritation of the gastrointestinal tract
     ." []
synonym: "anti-ulcer agent" RELATED [ChEBI:]
synonym: "anti-ulcer drugs" RELATED [ChEBI:]
synonym: "anti-ulcer agents" RELATED [ChEBI:]
is_a: CHEBI:23888
```

Listing 2: An entry for a chemical role — anti-ulcer drug — within ChEBI

1.4 RESEARCH AIMS

We would like to extend Chemical Entities of Biological Interest (ChEBI) to include both new edges of the sort currently added, which should make the ontology more *comprehensive*, and new kinds of edges — both within the ontology and between nodes in the ontology and nodes in other biomedical ontologies covering different domains — which should make the ontology *richer*.

ChEBI has not previously employed text-mining to support its enrichment. Given that it is an expert-curated resource, the approach chosen is to generate hypotheses which can then be fed into the existing curation pipeline as suggestions for the curators to add, rather than attempt to convert ChEBI into an automatically machine-curated ontology, which would risk diminishing the reliability of the resource.

We want to make suggestions which are:

1. high precision. This will help make good use of the curators' time by reducing the number of unproductive suggestions that they must

read and discard. In contrast, recall is not paramount, so long as the curators are not likely to exhaust the pool of suggestions.

2. mapped to entities which already have entries within the ontology. The scope of the exercise is to provide edges between existing nodes; providing a mapping will allow detection of edges that already exist within the ontology, and obviate the need for curators to identify the relevant nodes manually.

3. based on a variety of semantic relations if possible - minimally `is_a` and `has_role`. These edges make up the majority of edges within ChEBI, and hence their identification accounts for a large proportion of curatorial time. Providing other relations linking to other OBO ontologies may assist future expansion of ChEBI.

4. general. We would like not to have to train separately for each possible property of a chemical, as do Sun et al.(2009)[101].

5. linked to literature. This serves two purposes. One is that providing the text from which a suggestion has been extracted allows a curator to form a judgement speedily on the accuracy of the suggestion. The other is that the curation protocol includes documenting the sources of assertions within the ontology — providing these sources lessens the need to search for supporting citations.

6. ranked by confidence/importance/urgency of inclusion. It is highly desirable that these hypotheses be *ranked* so that curators can make best use of their time by considering the most likely hypotheses first; it is also desirable that hypotheses be *filtered* so that those that are already present in ChEBI (or able to be deduced from it) are excluded. It would also be helpful if we can have some kind of guide to how *informative* (non-obvious) a hypothesis is.

1.5 SUMMARY

This dissertation will deal with the intersection of the domains of biology and chemistry; looking at small molecules in a biological context, starting from the framework of an existing ontology, looking at ways to enrich the ontology by increasing the density of edges between existing nodes (without adding new nodes or new *types* of edges), and then extending it by examining the breadth of the expressions used to connect chemicals to other kinds of biological concepts.

Finally we look at ways that we might deduce connections between categories of chemicals where the connections have not been extracted from the literature in an explicit form. Examples include [*if a chemical* `potentiates` vasoconstriction *then it* `is_a risk factor-for` atherosclerosis], [*if a chemical* `is_a` fluorochrome *then it* `stains` DNA], and [*if a chemical* `is_a contaminant-of` food *then it* `is_a` toxin].

Chapter 2 covers some of the technology that has been used for bridging the gap beween the written word and a structured representation of knowledge, and Chapter 3 will explain in more detail the direction of this project in light of this previous work.

2

BACKGROUND

In this summary of literature, we shall look at the various steps that have been taken to bridge the gap between unstructured information, specifically the biomedical literature (although some of the relevant work has been on general-domain text or on other specific domains such as engineering), and structured information, specifically ontologies.

Most of the work has followed a pattern of i) identify entities, ii) identify relations between the entities, and iii) draw further conclusions from the relations — though this is not an invariant process, and especially some of the statistical approaches may analyse this approach differently. There are a variety of approaches that have been taken to detect relations, and these have been informed by the type of text being worked with, the resources available, the type of relations sought, and the uses for which the relations are intended.

The type of text being dealt with affects the approaches taken in a number of ways. Text from within a particular field (*domain*) may require tools that have been trained for that particular domain — for example syntactic parsers perform more accurately on biomedical text when they have been trained on such text[50] — and some types of text will tend to cause problems for non-adapted tools through unexpected punctuation use or capitalisation[54]. Text-specific or domain-specific vocabulary may well require specialised treatment such as disambiguation[100], and if lexical resources such as dictionaries or thesauruses are used, these may well not contain many of the terms that are important for the text in question. Most of the discussion below will focus on either *general-domain* text, which in practice is often based on such sources as news reports, or text specific to the field of biology and medicine.

The type of relations sought also affects the choice of approach. In some cases we can identify in advance which entities might have certain relations occurring between them — for instance we know that the *is-employee-of* relation occurs between a person and an organization, and for both of these types of entity* there exist many reliable Named Entity Recognition (NER)

* These, along with Location, make up the Entity Named Expression (ENAMEX) categories, which have been the subject of extensive work[75]

systems[75]. The types of relations between entities are relevant too. For some, we can form rules to help narrow down the plausible relations. For example, we can have a rule that states, for the relation [PERSON *dies-on* DATE] that a person can die only once*.

Another consideration is the methods of validation available. Some approaches require a *training corpus* — a body of text that has all the relations that are sought indicated (*marked up*). For some tasks, such corpora are freely available — this is particularly the case for well-studied problems. For other tasks, these approaches require that marked-up texts be compiled, either manually or automatically. The nature of the relations and the resources available (manual annotation requires a significant input of expert time and is thus an expensive procedure) govern which, if either, of these approaches are practical. Other approaches, while not requiring such a corpus, do require that we can, for at least a subset of the results, identify whether they are correct or not. Again, this can be done by human experts, or with an existing (partially complete) set of known correct and incorrect relations[68].

WordNet[67;66] is a lexical database for English, covering nouns, verbs, adjectives, and adverbs. It contains a diverse set of relations between words — including hypernymy and synonymy, and also such relations as meronymy (X has-part Y) and holonymy (X is-part-of Y). It has been widely used as a lexical resource for general English text, and for training and validation for tasks that seek to discover the kind of relations that it includes.

The uses that relations will be put to affect the techniques that are appropriate for their discovery. In general in such tasks, the quality of results is assessed in terms of *Precision* and *Recall*. Precision is calculated as the proportion of discovered items that are true; Recall is the proportion of true items that are discovered.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Which of these we wish to optimize for depends on whether the eventual use of the results is more sensitive to false positives or false negatives. Assessing recall is not possible for tasks where an estimate of the number of relations that exist in the text cannot be made.

A common metric used to evaluate the performance of systems is *F-measure* (also known as F_1 measure). Since many systems can trade off preci-

* This example is based on Aone and Santacruz(2000)[9] who use a rule that if two events involve the same person dying, they must be the same event

sion and recall, this measure, calculated as the harmonic mean of Precision and Recall

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + recall}$$

allows a single-number comparison of the performance of a system. Since some requirements place more importance on precision than recall (or vice versa), a higher value of the F-measure does not necessarily correspond to a more useful system, and there are variant (and less frequently-used) measures that allow different weightings for precision and recall to be specified. It is also important to note that different protocols for assessing precision and recall can make direct comparison of figures difficult — in particular treatment of partial or questionable matches. Nevertheless, F-measure is widely used and can be a helpful guide.

We also wish to consider whether the eventual use of the results requires transparency in the provenance of relations. For some purposes, a "black box" approach may be acceptable; for others it may be important to be able to track assertions back and identify what combination of text and algorithm gave rise to each relation.

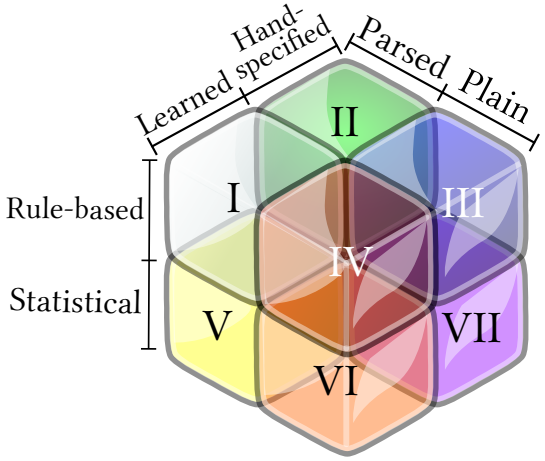


Figure 4: Three dimensions of approaches to relation extraction. X: approaches may use plain text or rely on syntactic parsing. Y: approaches may use qualitative rules or a statistical combination of biasing factors in order to detect relationships. Z: rules may be manually specified or machine-learned based on a training corpus.

Figure 4 illustrates possible combinations of three features of relation-extraction approaches. The three dimensions of approaches described in Figure 4 are a simplification, but helpful in comparing different systems for extraction of hypernyms and other relations. We should note that the "plain text/parsed text" axis in particular is not so dichotomous as it may

appear; systems may use *Part of speech (POS) tagging*, in which words are classified as nouns, verbs, prepositions, and so on, generally with indications of proper nouns and distinctions between singular and plural forms. They may also use "shallow" parsing, in which some phrase structures are identified [77]. "Deep" parsing identifies a tree of phrase structures, and possibly also some *dependency* structures, in which a graph structure is generated with its edges representing syntactic relationships between the words within a sentence. These are illustrated in Figure 5.

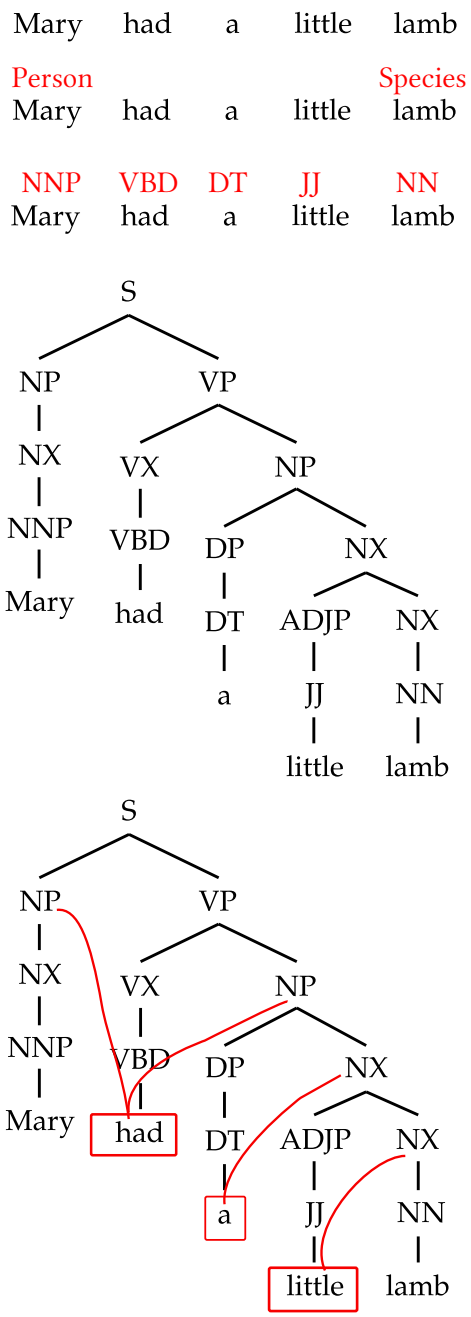


Figure 5: From top to bottom: Plain text, Named Entity Recognition, Part of speech tagging, phrase structure parse tree, and predicate-argument information for the sentence "*Mary had a little lamb*". Structures are as provided by Enju. Abbreviations are in Penn Treebank format[65]: NNP = Proper noun, singular; VBD = Verb, past tense; DT = Determiner; JJ = Adjective; NN = Noun, singular or mass. Red lines indicate the arguments of the boxed predicates — for instance the verb "*had*" has two arguments: the noun phrase that includes only the word "*Mary*", and the noun phrase that includes the words "*a, little, lamb*".

2.1 STATISTICAL METHODS

In an early example of statistical approaches to relation extraction, Church and Hanks(1990)[28] use statistical methods to identify, in general-domain text* the lexical environment that different words inhabitAll studies are with English language text unless otherwise specified. The motivation for this approach is to aid in disambiguation (for instance, in the results of optical character recognition), and to assist lexicography. They use an information-theoretical model (mutual information) to quantify the strength of association between neighbouring words and identify statistically-significant collocations — sets of terms that occur together more frequently than expected based on their overall frequency. Shallow parsing is used to identify parts of speech and the technique is applied to a set of subject-verb-object triples.

The approach of looking for co-occurring words has been applied to the detection of compound words and words that may be related *semantically* (in terms of their meanings) by Hamon and Nazarenko(2001)[49] (French language engineering domain). Turney(2001)[105] applies mutual information measures to the task of solving "Which of these words is the closest synonym for X"–style problems (based on tests aimed at learners of English), looking for frequencies of co-occurrence not by analysing texts directly, but by using search engine result counts. Smadja(1993)[97] also looks for collocations — arbitrary and recurrent word combinations — and co-occurrences, focussing especially on the preferences of certain verbs and nouns for particular prepositions.

A slightly different approach for identifying possible synonyms is proposed by Deerwester et al.(1990)[39]. The motivation behind the technique described, Latent Semantic Indexing, is indexing documents in a way that is robust to synonymy, by producing a matrix of terms in documents and then reducing its dimensionality. Cosine similarity, a measure of the similarity of the vectors that represent each term, can then be used to extract potentially synonymous pairs (pairs whose members tend to co-occur with the same set of other terms).

For problems with a wider scope than the identification of synonyms, techniques such as those used by Widdows and Dorow(2002)[111] identify not terms with the *same* meaning, but those with *related* meanings. Starting with "seed words", terms that share the most similar lexical environments are extracted, and these are then used as the basis of a cluster — terms can

* ,

only be admitted to the cluster if they share links with all the member terms. Use is made of WordNet as a resource for evaluating results.

In a small-scale proof of concept, Alfonseca and Manandhar(2002)[7] extend the use of WordNet from an evaluation tool to a target for enrichment. They use the text of JRR Tolkien's *The Lord of the Rings* as a source of text that is similar to general-domain language, but which has some specialist vocabulary; terms such as *hobbit* and *Mordor* are then automatically placed within WordNet in locations derived from the similarity of the environment ("topic signatures"[2]) of the new terms and the existing nodes within a carefully selected nine-node subgraph of WordNet.

Riloff and Shepherd(1997)[86] also extend groups for which a few exemplars are given*. They do this by noting which nouns occur more frequently in the immediate vicinity of other nouns within the group, on the basis that members of the same categories often co-occur in lists ("*Lions, tigers and bears*"). Their focus is on finding words that are thematically connected as well as strictly hypernyms.

2.2 SYMBOLIC METHODS

So far we have covered similarity measures and their applications. These have been statistically-based and aimed at finding either synonyms or terms that are otherwise somehow interchangeable. For other relation types, other techniques have tried to interpret the text symbolically to extract relations of various degrees of complexity.

An early and influential work in this area is Hearst(1992)[51]. This work seeks to identify `is_a` relations between Noun Phrases (NPs) in text by exploiting certain Lexicosyntactic Patterns (LSPs), which have become known as Hearst Patterns. A LSP is a specification of syntactic structure and text tokens that can be matched to pieces of text. The patterns used in this case include

```
NP such as {NP, {, NP}* (and|or)} NP
```

and

```
NP {, NP}* {,} or other NP
```

In these examples, `NP` represents any noun phrase. For example, by applying the second of these patterns to the phrase "*Bruises, wounds, broken bones or*

* For example, [*airplane, car, jeep, plane truck*] is a set that is designed to be extended with the addition of other vehicles

other injuries[...]", the rules *hyponym("bruise", "injury")*, *hyponym("wound", "injury")*, and *hyponym("broken bone", "injury")* are deduced. These results are validated by their agreement with WordNet, and the author characterises the results as being of high quality, although no numerical assessment of precision was performed. The approach yielded 330 hypernym/hyponym pairs from the text of *Grolier's American Academic Encyclopedia*.

Caraballo(1999)[23] uses a similar approach, using patterns of conjunction (*X, Y and Z*) and a cosine distance metric to create clusters of (single-word) nouns and then select the single best-attested hypernym based on the pattern of apposition: (*X, a Y*) with Y taking values such as "vehicle" or "clothing" for each cluster, working on the basis that each cluster has only one good-fit hypernym. This approach attains a higher yield than Hearst(1992)[51], but does not have high precision, with the top suggested hyponym for a given hypernym assessed as correct 33% of the time.

Cederberg and Widdows(2003)[25] use Latent Semantic Analysis (LSA) as an error reduction method, improving the precision of Caraballo's technique by discarding cluster members that are insufficiently typical of the set.

A move from the general to the biomedical domain requires varying techniques and resources. Rindflesch and Fiszman(2003)[90] use the Unified Medical Language System (UMLS) as a framework providing synonyms and semantic categories, and use appositive and copular (involving the verb "to be": *X is a Y*) patterns to extract hypernyms. A similar approach is applied to drug names from DrugBank by Kolářik et al.(2007)[60]. Both of these approaches use shallow-parsing of text to identify phrase structure.

In what might be described as a hybrid approach, Gurulingappa et al.(2009)[47] use Hearst Patterns on plain text to extract drugs that are classified as cardiovascular agents, and then use statistical measures of co-occurrence to identify potential properties of the drugs from phrases found near the drug names.

A slightly different approach is not to manually specify the LSPs but to design a framework in which they can automatically be derived. There are two main angles from which this can be tackled — using an unannotated corpus in conjunction with a reliable source of hypernyms, or using a "gold standard" corpus in which all the hypernymic relations have been manually annotated.

Morin and Jacquemin(2003)[72], working with unparsed French-language agriculture-domain text, take the first approach, looking for strings that connect pre-defined hyponym-hypernym pairs. They then identify the longest common substrings of these strings, to form a generalisation from examples

such as "*[...] analyse foliaire de quatre espéces ligneuses (chêne, frêne, lierre et cornouiller) dans l'ensemble des sites étudiés*" to patterns such as

```
{deux|trois...|2|3|4...} NP1 ( LIST2 )
```

.

Snow et al.(2005)[99], in general domain text, apply a similar procedure over a dependency-parsed representation of the corpus, and then make use of an annotated corpus in which all pairs of nouns have been labelled to develop a machine-learning (logistic regression-based) solution that can classify a path between two nouns as hypernymic or not, which provides a substantial improvement in yield and precision over Hearst Patterns.

2.3 NON-HYPERNYMIC RELATIONS

For relations other than hypernymic ones, as used in ontologies, some of the same techniques are usable[53]. Agichtein and Gravano(2000)[1] identify fixed relations between named entities such as ORGANIZATION *has headquarters in* LOCATION. For this, they use a Named Entity Recognition (NER) system which identifies in the text all instances of organizations and places, so that the strings between them can be classified. A set of known-true locations of headquarters is used to train the algorithm.

Aone and Santacruz(2000)[9] identify verb-based relations between the semantic types Person, Location, Organization, and Artifact; verbs and their arguments are mapped onto events (such as *Kills*, *Buys*) and relations (such as *Parent of*) based on manually-defined patterns.

Mintz et al.(2009)[68] use a large set of exemplars to extract from a very large corpus sentences that contain the entity pairs. These sentences are parsed to obtain a dependency structure. Combinations of syntactic patterns that imply the relations in question are automatically derived; a relation may not be derivable from any one sentence but still be derivable from a set of two or more.

Many of the more recent approaches use machine learning methods over parsed text to identify relations. Zelenko et al.(2003)[116] identify "person-affiliation" and "organization-location" relations within sentences from news text (restricted to those sentences that parsed correctly), obtaining an F-measure of 86.8% with use of a SVM-based classifier, by means of manually-curating a set of examples.

Zelenko et al.(2003)[116] and Miwa et al.(2009)[69] describe various methods by which parsed sentences can be converted into feature sets for such classification, using SVM kernels that analyse subsets of parse trees rather than sequences of tokens.

Within the biological domain too, non-hypernymic relationships can be extracted. The existence, especially for protein-protein and gene-regulation interactions, of such annotated biomedical corpora as Bioinfer[83], GENIA[59], and GREC[104], has enabled the development of a variety of systems for recognizing a set of biological events such as phosphorylation, binding, and gene expression; the precise set of events extractable varies with the corpus.[8]

Miyao et al.(2006)[71] is an example based on the GENIA corpus, classifying events into eighteen categories taken from GENIA's ontology of types of relation *. Which of these relations any NLP system sets out to detect varies: Giles and Wren(2008)[45] takes a more minimalist approach, classifying relations only by direction (whether X acts on Y or Y acts on X) and by whether the relation is stimulatory or inhibitory. Rinaldi et al.(2006)[87] and Villaverde et al.(2009)[107] use an LSP-type approach over parsed text, extracting a set of relations such as *activate, bind, block, regulate, control, express*.

Looking to an even more specific field, that of chemicals and drugs within biomedical literature, there are few appropriate corpora, although there have been some very recent efforts such as by Rinaldi et al.(2012)[88] to automatically curate such resources. Giles and Wren(2008)[45], training a Support Vector Machine (SVM) on GENIA, cover a chemical, caffeine, as one of their examples, suggesting that for tasks such as identifying the presence or absence of a regulatory relationship, a non-chemical-specific corpus may be useful. Most others have taken the approach of using manually-specified patterns; Rindflesch and Fiszman(2003)[90] use such patterns in identifying relations such as [*X* TREATS *Y*] between drugs and diseases, and Rindflesch et al.(2000)[89] cover interactions between genes, cells, and drugs. Gurulingappa et al.(2012)[48] describe the development of a corpus annotated specifically for events of the type [Drug causes adverse-effect], which is again a very specific relation type.

Sun et al.(2009)[101], looking for information on chemicals which may affect cancer risk, take the approach of training a classifier, using features based on keywords within the text to assign documents whose subject is the chemical under scrutiny into any of a number of classes each representing a mode of action, type of evidence, and so on. This approach performs

* http://www.nactem.ac.uk/tsujii/aNT/event.html

well with a limited set of classes (48 in total, though much of the work was only performed with the 37 classes that were well-represented in the training corpus). To extend the number of classes would need a much larger training set.

2.4 SUPPORTING ONTOLOGY CURATION

Various approaches have been taken to develop automatic systems to assist in manual curation of ontologies (and, relatedly and more commonly [79], annotation of papers which may then be used as a source of data for ontology curation). These are frequently referred to as *semi-automatic* approaches [43].

PaperBrowser [21;57], developed as part of the FlySlip project, is a system which is designed to aid the curators of the FlyBase *Drosophila* genomics resource* in their reading of an article, by highlighting named entities (protein/gene names) and providing partial anaphora resolution, thus allowing rapid identification of all mentions of a particular entity (whether by name or not) throughout the paper. PaperBrowser was found to provide a substantial speed advantage to curators; an efficiency improvement of 58% was found following a detailed evaluation.

Many of the other studies aimed at producing semi-automated ontology curation tools did not provide details of any assessment their reception by the curators themselves or the effect on the curatorial workload. However, Alex et al.(2008) [5], looking at the curation of a database of yeast protein-protein interactions, also make an attempt to establish an upper limit for the speedup that such a system could provide, by supplying curators with a gold-standard set of known-good interactions; they find that, for this task, a reduction in curation time of ⅓ is possible under these conditions, and conclude that a perfectly-performing NLP pipeline would have similar effects. They also establish a subjective preference among curators for a high-precision suggestion set over a high-recall one.

With an eye to identifying relationships rather than entities, Reinberger and Spyns(2004) [84], working within the medical domain, describe a clustering approach based on shallow-parsed text, focussed on identifying terms that occur in similar contexts. The exact relationships that will link the terms in an ontology is left to the expert readers rather than being suggested by the software.

Fortuna et al.(2006) [43] describe OntoGen, an interactive system to suggest topics for inclusion in a topic ontology — in which the main relation type is

* http://flybase.org/

`subconcept_of` — based on latent semantic indexing and k-means clustering. In this case, relations were predicted based on the data, though the relation type is somewhat semantically loose.

The *SciBorg* system, described within Copestake et al.(2006)[32] was developed for the chemical domain but with the aim that it be applicable more widely. It is based around RMRS[31], a framework that allows a systematic encoding of the semantics of a natural language statement. One of the anticipated applications of the project was the semi-automatic extension of ontologies.

The approach of annotation systems is investigated by Winnenburg et al.(2008)[113]. Existing ontologies are used to identify and resolve terms within documents, allowing terms to be linked to entries in ontologies or lexicons; curators can then use these documents as they would unlabelled documents, but with greater ease.

Text2Onto[29]* extracts a variety of candidate relationship types, which are then presented to a user via a graphical user interface. As the user provides feedback, the Text2Onto engine adjusts its probabilistic model accordingly, which helps improve the accuracy of future candidate terms, reducing the workload of the user.

Van Auken et al.(2009)[106] describe a semi-automated system based on Textpresso† that is specifically aimed at assigning GO terms of the *Cellular Component* subtype to proteins from the nematode *Caenorhabdtis elegans*; that is, classifying where a protein is localized within a cell.

The DOG4DAG system, described by Wächter and Schroeder(2010)[108] is a plugin to the popular OBO-Edit‡. DOG4DAG provides resources for a curator adding terms to a new or existing ontology. It uses a mixture of a dictionary-based approach — querying other OBO ontologies to find nodes whose descriptions match the term in question — and lexicosyntactic patterns (variations on Hearst Patterns) to identify potential definitions for the term and potential hypernyms. While it uses part-of-speech tagging and some simple heuristics for noun phrase detection, it does not make use of full syntactic parsing, though it does also use statistical methods to identify collocations within text that are likely to represent potential ontology terms.

More recently, Gobeill et al.(2013)[46] describe a comparison, at different times, of machine learning (*k*-nearest neighbours)-based and dictionary-based approaches to identifying candidate GO terms to assign to a protein with the aim of assisting curators expanding the GOA ontology. In practice

* Not to be confused with its predecessor *TextToOnto*.
† `http://www.textpresso.org`
‡ `http://www.oboedit.org`

this approach was assigning classifications to a paper that had been identified as concerning the protein. The findings were that as the database of known properties of proteins had grown, the machine learning approach had rapidly exceeded the dictionary-based approach in effectiveness, as measured by the system's ability to rediscover known-correct properties of proteins in GOA. Even though the same period saw a threefold increase in the number of synonyms in GO, this had only a very modest effect on the performance of the dictionary-based approach. They conclude that *"machine learning approaches are now nearly able to reproduce the quality of GO terms assignment as performed by trained human curators"*. However, systems such as the machine learning approach they describe do not tend to provide direct textual evidence for assignments, which is to be desired in the curation of many ontologies.

3

OVERVIEW OF THIS THESIS

In the last chapter, we saw some of the approaches that have been used within the fields of hypernym and relation extraction. In this thesis I will be describing the development of a system to identify these relations, specific to the ChEBI ontology, with the aim of aiding curation of the ontology. This chapter provides an overview to the system and how it was designed.

3.1 APPROACH

As discussed in Section 1.4, we would like to generate suggestions for addition that meet a number of criteria - precision, compatibility with the existing ontology, generality, traceability with respect to the literature from which they are derived, and ranking to make best use of curatorial time.

It is therefore highly desirable to have a system which can present to curators a set of suggestions of properties of a given chemical species which are well attested in the literature, along with links to the supporting textual evidence. Because these suggestions are to be presented to humans directly, they should be of high enough quality that they do not waste curatorial time — ideally a very small proportion of the suggestions should need rejecting. While we would obviously wish for such a system to have a high yield, this is not nearly as essential as high precision, since the rate of additions to the database is limited primarily by the availability of curatorial time. Since we would like explicit links from the literature to the suggestions, we might prefer a text-mining approach, where the suggestions are sought directly from a corpus of biological literature, to one that infers properties of chemicals from other sources.

Figure 4 described three dimensions of variation: Parsed/plain; Automatically derived/manually specified; Rule-based/statistical. Which of these is most appropriate?

- Parsed text will allow more complicated patterns to be examined, and the ability to use phrase structure which parsing confers will be important in identifying multi-word arguments. (cf. work in Villaverde et al.(2009)[107] which uses phrase structure and part of speech tag-

ging although it does not use any predicate-argument structures in identifying patterns). Plain-text-based LSPs will not properly identify phrases as in Figure 15. Gurulingappa et al.(2009)[47], applying Hearst Patterns to the task of identifying hypernyms within the restricted domain of drugs listed in the ATC, recommends as further work the exploitation of deeper analysis of sentence structure in order to extract more information.

- In the absence of a large training corpus, hand-specified is much easier to design. Such a corpus is not available for this task — protein- and gene- focussed work (e.g. Ananiadou et al.(2010)[8]) uses corpora such as GREC, GENIA, and BIOInfer, which do not have good coverage of chemicals. We would need something that not only had markup of chemical entities, but also of their properties.

- A rule-based approach allows us to add new patterns for different relationships; additionally, hand-specifying statistical patterns is not practical. If we do not have good corpora to assess patterns, we can perhaps have more confidence in patterns for which we can readily inspect the workings (although it remains important to numerically assess the output of the system as far as possible). While we would like to learn LSPs as in Snowball[1], the task of identifying company headquarters, where there is a definitive list that can be used to generate a supervised training system, is not comparable to identifying the open-ended relations between chemicals and their properties, which are hard to define, very context-dependent, and where even having a definitive partial set (such as ChEBI) does not allow us to automatically identify negative training examples.

We therefore settle on the combination of parsed text and hand-specified rule-based extraction of relations.

Turning now to the nature of the parsing, using an NER preprocessing stage has certain advantages:

- It reduces the quantity of subsequent processing and storage needed, as we can ignore sentences that do not contain any mentions of chemical entities

- It improves the reliability of parsing, which is not otherwise very robust to some punctuation characters that are common within chemical names.

- Since we are using patterns corresponding to common syntactic structures, and many chemicals have synonyms which are also common words (*lead* as a verb, *NO*, *As*, etc.), it will reduce the rate of false positives if we can use a context-dependent NER step to only identify those tokens as chemical when we have more evidence that they are indeed chemical entities.

The NER techniques employed, described in Chapter 4 are reasonably similar in nature and in performance to those in OSCAR3 [34], though since many of the functions performed by that package were not required, a less versatile but somewhat faster implementation was developed. We use the Enju parser, as described in Chapter 4, and a set of hand-specified lexicosyntactic patterns, described in Chapter 5.

Figure 6 shows a graphical representation of the adopted workflow.

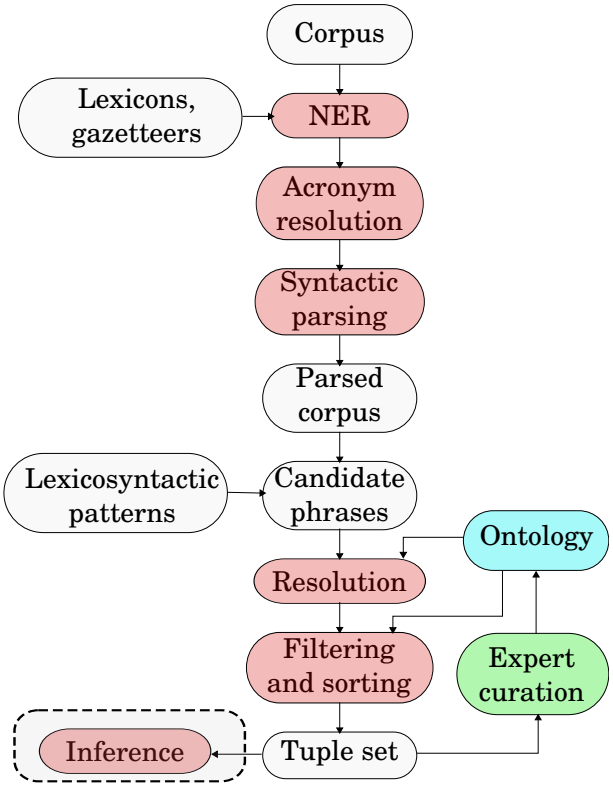


Figure 6: Summary of workflow. Grey and blue items represent resources; red items represent processes. The curatorial action is highlighted in green. The optional inference section is shown within a dashed line.

3.2 ASSESSMENT OF HYPERNYMS

Having decided on this approach, it is important to evaluate the results of detection of `is_a` and `has_property` relationships. These were assessed by

two of the ChEBI curators, who judged whether the identified relationships were true according to ChEBI inclusion criteria, this being the relevant measure for an exercise intended to enrich the ontology.

As discussed above, supporting manual curation requires high-precision suggestions if we are not to waste the curators' time, and there is a balance to be struck between this and producing a tiny data set containing only the suggestions about which we are absolutely certain. One of the features that it would be desirable to have is a method of filtering out hypernyms of which we are less certain, to the degree required for the application in question. I identify and investigate some of the various features that affect the accuracy of hypernyms. These were combined using a logistic-regression-based classifier to enable us to assign a confidence score to each hypernym. Discarding suggestions below a given level of confidence can adjust our precision in a continuous manner. This is discussed in Chapter 5

3.3 EXTENSION TO NON-HYPERNYMIC RELATIONS

An important benefit of the use of LSPs is that — unlike purely statistical association techniques — they are tractable to expansion from hypernymic to non-hypernymic relationships. While a chemical's hypernym will generally be another ChEBI-type concept, chemicals can interact, within a biological context, with a wide range of different types of entity. In Chapter 6, I use three other Open Biological Ontologies (OBO) ontologies as examples, covering the domains of diseases, proteins, and the tripartite GO categories of Cellular Component, Biological Process, and Molecular Function.

Three productive patterns are implemented (and the technology would be easily extensible to implement others). The patterns chosen are based on an examination of the hypernyms that were resolved insufficiently specifically in Chapter 5.

Given that ChEBI does not currently incorporate links to these other ontologies, and there is not an existing framework into which edges can easily be added, the first step must be to extract a broad sweep of relations, to obtain a survey of the kinds of concepts that are being discussed and asserted within the biological literature.

Whereas many projects, for instance Rinaldi et al.(2006)[87], have used a limited set of verbs chosen to correspond with known biological events (activation, inhibition, secretion, etc.), I have taken the converse approach of extracting as wide a set as possible, both to provide a potential basis for

extending inter-ontology links, and as a knowledge-base to be mined for implicit associations. For this reason I refer in general to "relations" rather than "events" as I do not wish to restrict myself to biological events like inhibition or secretion, but to mine as large as possible a set of ways in which entities can relate to each other. The use of a parser that yields predicate-argument relations enables the extraction of a cross-section of the interactions that are found between molecules and biological entities.

3.4 INFERENCE OF RELATIONS

This broad trawl of relation types requires post-processing, and one possibility for this is opened up by the variety of semantic types used. Each relation is profiled based on the spread of semantic types that it applies to. This can be used as a noise filter or to help select relations with similar meanings. Using a Pointwise Mutual Information (PMI) measure to identify similar relations also allows some automatic conflation of synonyms to help reduce sparsity in the data set.

Because the relation types are not defined in an existing ontology, we rely for validation on the agreement between the evaluators, which should indicate whether a given assertion is meaningful or not. Even working with limited definitions, they achieved a high level of similarity in their results, suggesting that there is a reasonably objective basis to their findings.

Beyond the explictly-stated relationships, we may well wish to generate suggestions for implicit relationships between properties, both to inform ChEBI development, and to provide hypotheses for biological research. While much research has concentrated on being able to characterize entities based on their properties, I have taken steps to characterize properties based on the entities that share those properties, going beyond the scope of Bodenreider et al.(2005)[18], and Bada and Hunter(2007)[13] in that I use a much wider set of properties discovered from the literature, as opposed to being restricted to the lexical forms of relations used within the ontologies themselves and the assignments within Gene Ontology Annotation (GOA). I demonstrate how different inference techniques for association rule discovery can be used to generate useful candidates for synonyms and for ontological relations. This is described in Chapter 7.

Also in Chapter 7, I describe a format of evaluation based around the "but-test" that utilises semantic oddness to elicit a judgement of biological typicality from human evaluators. This aims to identify which features are

thought of as applying in essence to a group, rather than applying strictly to every member of the group — we see a similar effect with bare plurals, so that "birds can fly" is in some sense valid, (penguins notwithstanding), and "mammals lay eggs" is not valid (platypuses notwithstanding).

Figure 7: A penguin, notwithstanding.

4

NAMED ENTITY RECOGNITION AND PARSING

This chapter describes my development of a system to detect occurrences ("mentions") of chemicals in scientific literature. The system is designed to examine each word in the text and predict whether it is likely to be part of a chemical name, based both on attributes of the word and on attributes of its neighbours within the sentence in which it is found. It is designed around a (third-party) "machine learning" tool for generalising from a set of examples rather than relying on rules supplied by me. It is implemented as a service task to identify chemicals in text prior to the detection of relations as described in Chapters 5 and 6.

4.1 BACKGROUND

NER is the process by which words* in a passage of text are identified as representing (in their meanings) examples of a particular class of object. Examples of categories of things that might constitute this class are names of people, places, companies or organizations; or within the biomedical domain, things such as proteins, genes, chemicals, cell lines, subcellular structures, and so on.

A simple approach to this task is a dictionary-based one - one starts with a list of names, and then classifies each token in the text according to whether it is present in the dictionary. This approach has the advantage of speed and simplicity, but tends not to produce very reliable results for most domains, due to its inability to take into account context and the prevalence (for most domains) of synonymy (multiple words which are spelt the same, such as *lead* "a chemical element, atomic number 82" vs *lead* "to guide or conduct") and polysemy (multiple senses for the same word, such as *gold* "a chemical element, atomic number 79" vs *gold* "money"). This also has the problem that a comprehensive dictionary is required, which is effectively impossible for most categories.

A more effective approach, given an appropriately annotated corpus, is to automatically learn the factors that indicate whether a given token is part of a chemical or not. A description of how this was perfomed follows.

* Strictly speaking, tokens; Section 4.2 explains in more detail.

4.2 DEVELOPMENT OF A NAMED ENTITY RECOGNITION SYSTEM

Tokenization is the process by which a string of characters — letters, numbers, spaces, and punctuation — (henceforth just "string") is grouped into smaller strings each of which corresponds to "those basic units which need not be decomposed in a subsequent processing"[110]. These normally correspond approximately (in English) to units that would be colloquially described as words, though it should be noted that (a) there is not a universally-accepted definition of what constitutes a word, and (b) for many tasks some tokens may be larger or smaller than a single word.

Considering the definition above from Webster and Kit(1992)[110], since we are attempting NER of chemical entities, we try to choose a token definition such that each chemical occupies one or more whole tokens. We wish to avoid the situation where the boundary of a chemical name falls within a token, so as to have no tokens that are "partially" chemical entities.

A simple and common approach to tokenization is to split a string up on spaces and punctuation such as commas, apostrophes, and hyphens. Unfortunately for the chemical domain this will tend to yield tokens that are somewhat smaller than desired, especially tending to break chemical names which often include all of the above. This is a problem because the more tokens a chemical name is broken into, the more judgements have to be made correctly to correctly classify all the tokens.

So, for instance, if *1,1,1-trichloro-2,2-di(4-chlorophenyl)ethane** is construed as a single token, in a simple system it must be either (correctly) identified as a chemical, or (incorrectly) identified as not-a-chemical. If, on the other hand, it is construed as the nineteen tokens "*1*", ",", "*1*", ",", "*1*", "-", "*trichloro*", "-", "2", ",", "2", "-", "*di*", "*(*", "*4*", "-", "*chlorophenyl*", "*)*", and "*ethane*", then all nineteen tokens must be classified correctly in order for the boundaries of the chemical to be identified, with nineteen decisions needing to be taken correctly.

To minimise the number of tokens per chemical entity, I chose to use the tokenizer module from the OSCAR3 package[33], which is designed to break apart chemical entities as little as possible. The tokenizer is based on a set of simple rules that are designed to identify non-alphanumeric characters that are used as part of a chemical name rather than as punctuation. These rules are described in Corbett et al.(2007)[35].

I then annotated each token with a set of features, including some boolean (either true or false) features (such as "begins with a digit"); string features

* IUPAC name for the pesticide DDT

such as letter-trigrams and -tetragrams, and stems as given by Porter's algorithm[82]; and numerical features representing the probability that the token represents a chemical entity based on a Bayesian model using the relative frequencies of the letter-ngrams within the token in ChEBI and PubChem[109], used as a large sample of chemical names; and the British National Corpus (BNC) (`http://www.natcorp.ox.ac.uk/cpr.xml?ID=reference`), used as a large sample of non-chemical text. All of these corpora were tokenised using the tokenizer described above. This is an simplification of the method of Corbett and Copestake(2008)[33], who use a letter-based Markov model for a similar purpose.

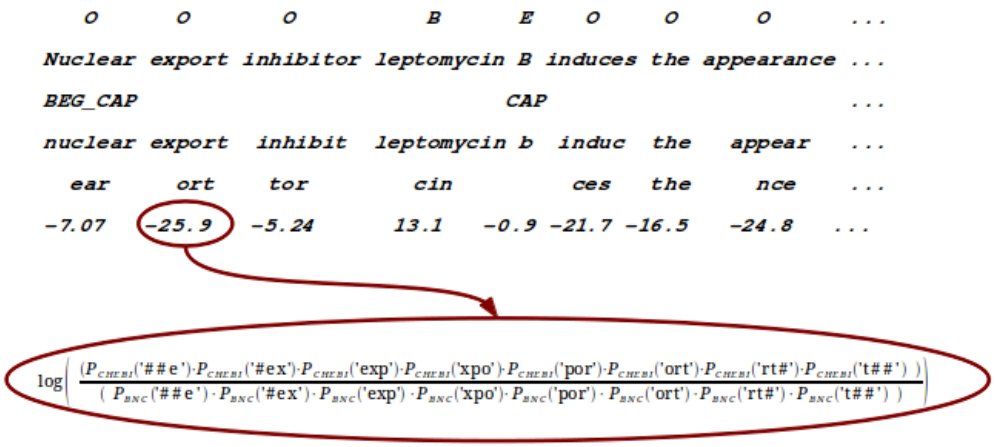


Figure 8: A sample of text annotated with a subset of NER features (raw text, capitalisation, stem, three-character-suffix, and estimate of probability that it constitutes a chemical name according to a letter trigram model and the ChEBI set of chemical names), and shown below labels indicating chemical entities. B indicates a token at the beginning of an entity; E at the end; O indicates a token not within a chemical entity.

Table 1 contains the full list of features used in the classifier. The Boolean features are largely based on those in Bikel et al.(1997)[15], Collier et al.(2000)[30], and Shen et al.(2003)[95]. While Shen et al.(2003)[95] pre-classify suffixes as describing semantic types, I use features for prefixes and suffixes, as well as letter trigrams and tetragrams to allow a classifier to learn such implications, as do Corbett and Copestake(2008)[33].

The SVM-HMM package[56], which implements support-vector-machine classification as applied to learning transition probabilities between labels in a sequence, was used to identify chemical entities based on the feature vectors of each token combined with the feature vectors of the tokens in the -2, -1, +1, and +2 positions (see Figure 9 for an illustration*). SVM-HMM was trained on the annotated corpus of Corbett *et al.*[35]. Abbreviations of

* The text used is the standard typesetter's dummy text *lorem ipsum*. An informal summary of the history of the text can be found at `http://www.straightdope.com/columns/read/2290`.

Feature	Type	Description
_BEGINNUM	Boolean	Token begins with a digit
_PERCENTAGE	Boolean	Token matches `/^\d[\d\.]*\%$/`
_INITCAP	Boolean	Token begins with [A-Z]
_ALLCAP	Boolean	Token matches `/^[A-Z\d]*[A-Z][A-Z\d]*$/`
_NONALPHANUM	Boolean	Token does not contain alphanumeric characters
_OPENBRAC	Boolean	Token begins with an openbracket (round, square or curly)
_CLOSEBRAC	Boolean	Token ends with an closebracket (round, square or curly)
_SINGLEDIGIT	Boolean	Token consists only of a single digit
_DOUBLEDIGIT	Bolean	Token consists only of two digits
_INTEGER	Boolean	Token matches `/^-?\d+$/`
_REAL	Boolean	Token matches `/^-?\d\.+\d+$/`
_ROMAN	Boolean	Token matches `/^[IVX]+$/`
_QUOTE	Boolean	Token matches `/^[\'\"\‘]+$/`
_STEM	String	Stem of the token, per Porter's algorithm[82]
_LC	String	Lowercase of token
_END3	String	Last three characters of token
_END2	String	Last two characters of token
_PRE3	String	First three characters of token
_PRE2	String	First two characters of token
_TRIGM	String array	Three-character subsequences of token
_TETGM	String array	Four-character subsequences of token
_GAZETTEER	Numeric	Number* of times the token appears in a tokenized gazetteer of chemical names
_TRI_SCORE	Numeric	Probability that token belongs to a PubChem chemical name, based on letter trigrams
_TET_SCORE	Numeric	Probability that token belongs to a PubChem chemical name, based on letter tetragrams
_TRI_SCORE_C	Numeric	Probability that token belongs to a ChEBI chemical name, based on letter trigrams
_TET_SCORE_C	Numeric	Probability that token belongs to a ChEBI chemical name, based on letter tetragrams

Table 1: Features applied to each token for NER

the more common *Long form(ABBREV)* type were identified according to the method of Schwartz [93] and attached to each article. Based on the principle of "one sense per discourse" [44], if a string was identified as representing a chemical entity, then all other instances of that string were similarly annotated. If the string was identified (anywhere within the same abstract) as having an abbreviation, then all instances of the short form were also annotated. So, for example, if we see the sentence *Composites with polypropylene (PP) and jute fiber were prepared by injection molding technique*, then having judged that a) *Polypropylene* is a chemical name, and b) *PP* is an abbreviation for *polypropylene*, all instances of either *PP* or *polypropylene* in that text will be marked as chemicals. It is important that this is limited to this text, since *PP* is widely used as an abbreviation for non-chemicals, such as *Protein Phosphatase* or *Physiopathology*.

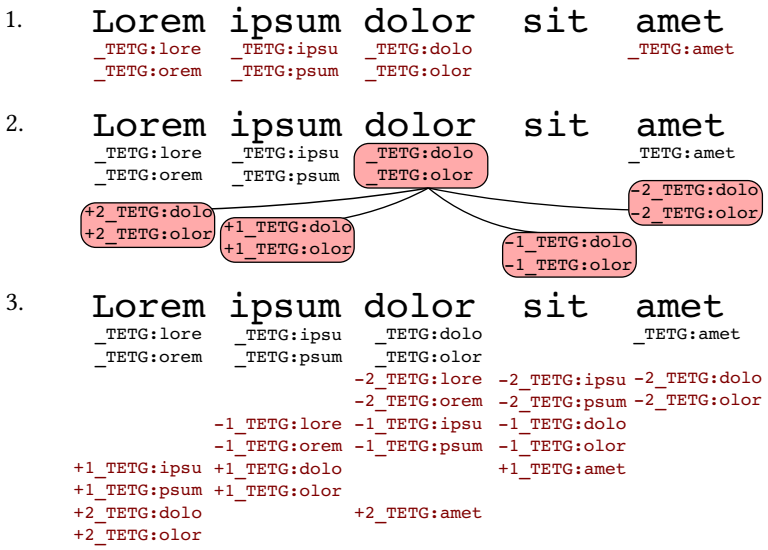


Figure 9: Here we see a simplified example of propagation of tetragrams between neighbouring tokens. 1: All tokens are labelled with their constituent letter tetragrams. 2: Each label is copied to the two preceding tokens and the two following tokens. 3: The resulting set of labels.

4.3 EVALUATION OF NER

In order to evaluate the effectiveness of NER with different amounts of training data, a range of trainingtest splits were used. The results are shown in Figure 10. A chemical entity was considered to be correctly identified if-and-only-if the NER system detected an entity with identical boundaries. Partial

matches (e.g. overlaps) were considered to be false positives. Precision is calculated as

$$\frac{true\ positives}{true\ positives + false\ positives}$$

and recall as

$$\frac{true\ positives}{true\ positives + false\ negatives}$$

. F is calculated as

$$2 \cdot \frac{precision \cdot recall}{precision + recall}$$

It can be seen from the graphthat with a large training set, precision in excess of 90% and recall over 80% could be achieved. It was not easily possible, using the SVM-HMM software package, to explicitly vary the trade-off between precision and recall.

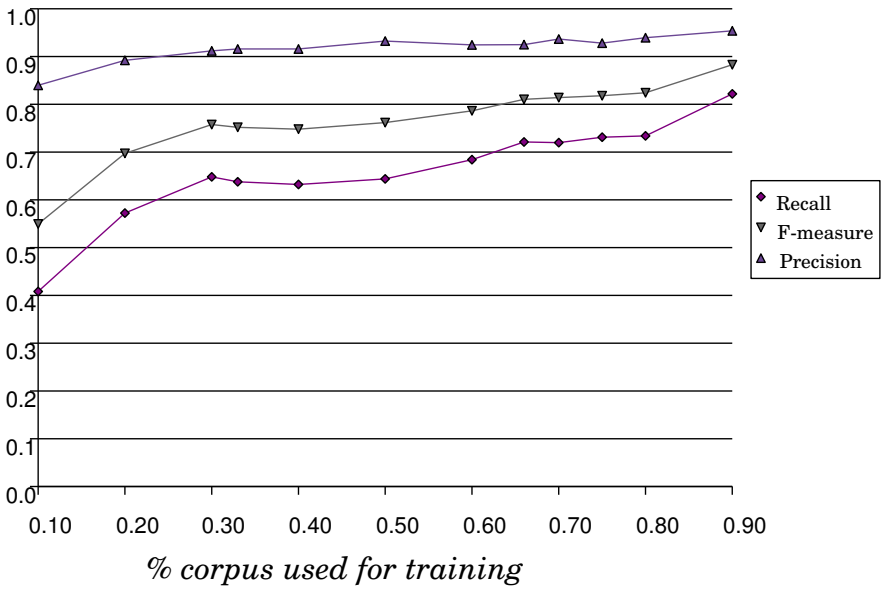


Figure 10: Precision, Recall, and F-measure for named entity recognition using differing proportions of the data set as training data

4.4 USING NER RESULTS FOR PREPROCESSING BEFORE PARSING

The *Enju* package[70] was used for syntactic parsing. It accepts plain text and returns as output an XML representation of the phrase structure of the sentence and the predicate-argument relations between tokens and phrases.

Enju has two text models – one for general-purpose English text, and the one used throughout this work, their "Genia model", which is optimized for biomedical text. However, Enju's tokenizer is a simple-pattern-based one. There are special uses of punctuation characters in chemical names (Figure 11 illustrates a chemical name containing a comma and a hyphen), and Enju — even using the Genia model — has a strong tendency to treat these as though they were normal punctuation. Accordingly, all non-alphanumeric characters (including spaces) within identified entities are replaced with underscore characters (Figure 12), leaving a single string which is more readily correctly parsed.

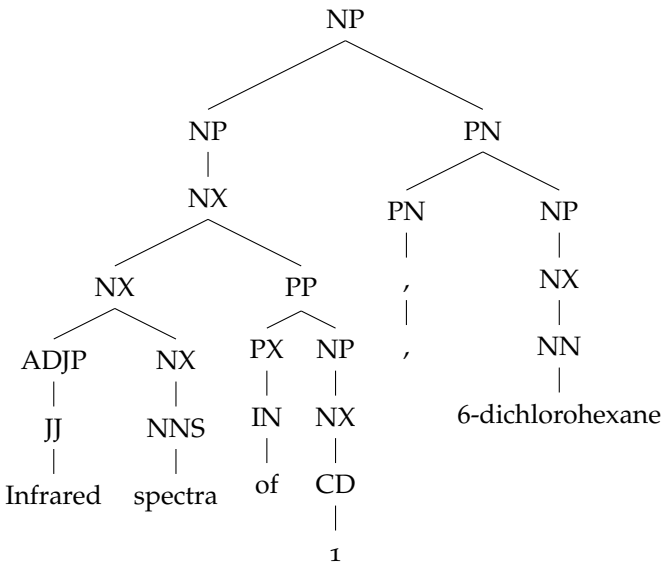


Figure 11: The result of parsing the raw text "*Infrared spectra of 1,6-dichlorohexane*".

After excluding those sentences that did not contain mentions of chemical entities, and those that (including some non-sentences such as "`II.`", as well as sentences not in English), 9,311,941 sentences (from 2,527,800 papers) remained for use in hypernym searching, as described in Chapter 5.

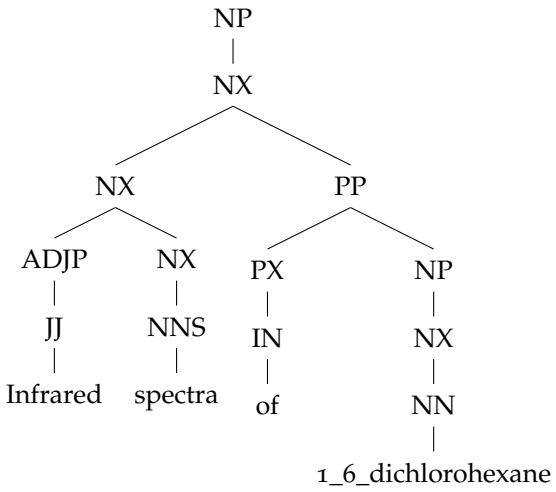


Figure 12: The result of parsing the text "*Infrared spectra of 1,6-dichlorohexane*" with underscores replacing non-alphanumeric characters within chemical entities.

5

IDENTIFICATION OF HYPERNYMS

In this chapter, I detail how biological literature can be processed so as to yield sets of is-a ("hypernym") relationships mapped onto an existing resource, ChEBI, which contains a large number of facts about chemicals, their properties, and the relationships between them. While ChEBI is hand-curated, this system allows suggestions of additional facts for inclusion to be made automatically.

5.1 BACKGROUND

Ontologies can be described as graphs possessing nodes and edges[17]. In ChEBI, the nodes can be divided into those representing chemical species and those representing roles or properties of the chemicals. The edges are overwhelmingly of the types `is_a` and `has_role` (Chemical `is_a` Chemical; Role `is_a Role`; Chemical `has_role` Role). It is these edges that we shall consider here, although ChEBI has other edge-types (such as `has_conjugate_acid` and `is_tautomer_of`).

When considering the linguistic realisations of `is_a` and `has_role` edges, we find that they are, by and large, spoken of indistinguishably. The same language is used to declare that *glutamate is an amino acid* (`is_a`) as *glutamate is a neurotransmitter* (`has_role`). Throughout this thesis, therefore, the two types of edge are treated interchangeably on the understanding that the knowledge of the types of the nodes will allow a curator to infer the proper type of edge for inclusion in the ontology:

Frogs are [has_property] green, and I'm [is_a] a frog, and that means I'm [has_property] green — Kermit the Frog (Joe Raposo & Jim Henson), 1970

Hyponym	Hypernym	Type
Chemical	Chemical	`is_a`
Chemical	Role	`has_role`

A Role is identified by having an `is_a` relationship (*sensu stricto*) with `CHEBI:50906` "role"; a Chemical is identified by having an `is_a` relationship (*sensu stricto*) with `CHEBI:24431` "chemical entity".

In the automatic detection of hypernymy relationships, there is a spectrum of approaches that have been taken. There is the high precision (most of the detected hypernyms are correct) but extremely low yield (applica-

tion to a given corpus does not produce a large number of detected hypernyms) approach of using the textual lexicosyntactic patterns described by Hearst(1992)[51], as discussed in Chapter 2. Hearst's approach has been varied by employing other patterns such as those automatically derived by Riloff(1996)[85] for identification of noun phrases or hyponyms[6].

There are also higher-yield but sometimes less precise methods that use statistical association between terms[16;63]. Statistical methods also have the drawback that it is not possible to link an assertion to a source text. The ability to do this is one of the requirements of ChEBI, where each entry in the database must be explicitly attested in the scientific literature.

There is something of a middle ground occupied by techniques that use less specific textual patterns. An example is the work of Snow et al.(2005)[99] who use WordNet to label hypernyms in a dependency-parsed corpus and then compare dependency paths linking hypernyms to those linking non-hypernyms, in order to retrieve a large family of paths that preferentially identify hypernyms rather than non-hypernyms. Such techniques require better training resources (in the case of Snow et al., utilising WordNet) than are available for this domain-specific problem.

Systems such as Snowball[1] have used bootstrapping based on identifying lexical patterns between the hyponym and the hypernym. This type of detection, however, relies on having a good named-entity recognition (NER) system for the entities (which are noun phrases of variable length) involved in the relationships. We lack such a system for the hypernyms in this case, although we can identify the hyponyms (see Chapter 4). This lack also presents a barrier to using systems relying on statistical association between chemicals and their descriptors – we would need to be able to reliably identify the hypernyms in free text.

The solution adopted is based on simple syntactic patterns that identify appositive (having a noun phrase defining a neighbouring noun phrase: *e.g.* "*X, a Y*") and copular (involving the verb *to be*: *e.g.* "*X is a Y*") relationships within the trees yielded by syntactic parsing.

By restricting the output of a system to sentences where both the hyponym and hypernym can be resolved to terms in a controlled vocabulary, and by additionally using a named entity recognition system that is sensitive to context, the number of false positives generated can be reduced enormously.

While the use of such patterns does not in itself mandate a NER step, there are several major advantages to performing such an operation. The first is that a fast NER step can allow us to discard sentences that do not

contain any mentions of chemical entities, providing a large speed gain for all subsequent steps.

The second is that a context-sensitive named-entity recognizer will add further confidence that the phrase designated as a chemical entity later in the process actually is so. There are many chemicals (including abbreviations) that are homonyms of reasonably common English words (*e.g.* "lead", "NO"). Additionally we wish to exclude chemicals that occur as a premodifier in compound protein names (*e.g.* "succinate dehydrogenase" is not the same as "succinate"), an important consideration when dealing with biological literature.

The third advantage is that the syntactic parsing of sentences containing chemical names is rather unreliable due to the prevalence of commas, hyphens, brackets, and so on inside such names. An NER step allows these characters to be removed within chemical names, and then restored after the parsing stage is complete.

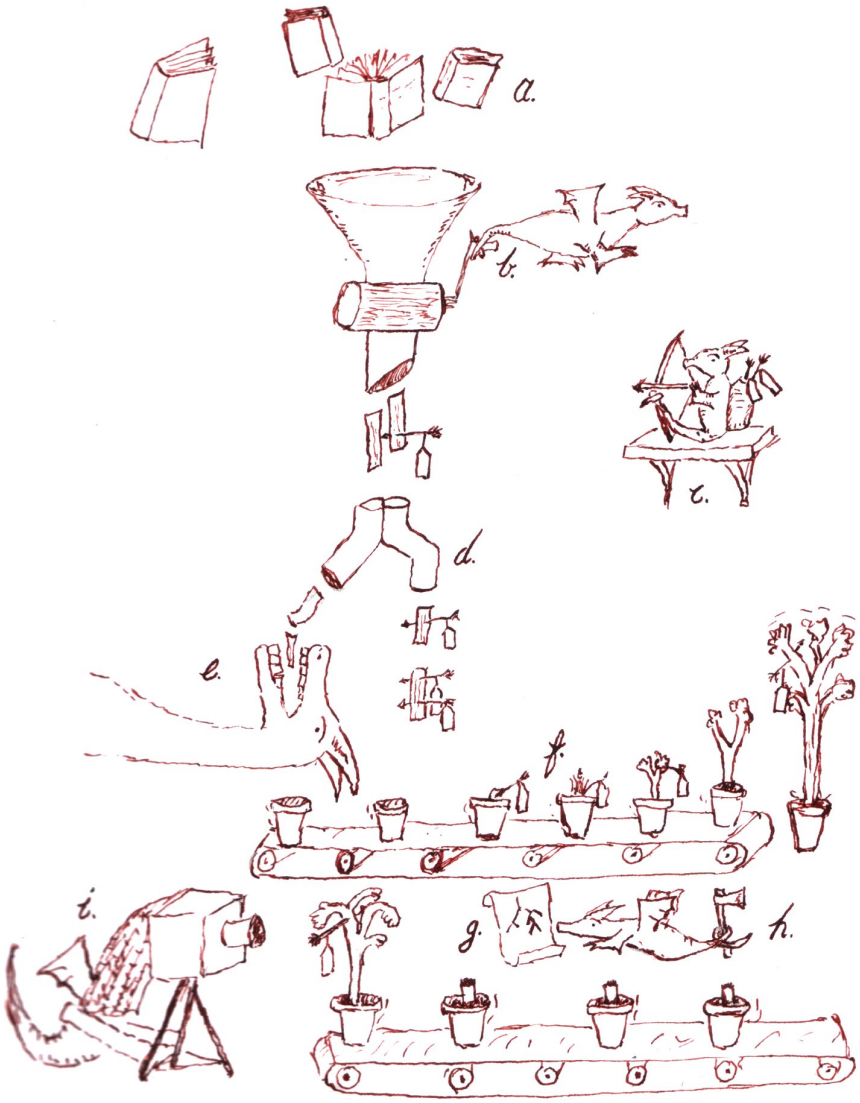


Figure 13: A partial depiction of the workflow for preparing text for relation detection. a) Literature is added at the top. b) Documents are reduced to sentences. c) A tagger identifies mentions of chemicals in the text. d) Sentences with mentions of chemicals are kept. e) Sentences without chemicals are discarded. f) A parser builds syntactic trees for the retained sentences. g) A set of hand-drawn patterns specifies which sentence-trees are interesting. h) The uninteresting sentences are thrown away. i) The interesting trees are recorded for examination.

5.2 METHODS

In total, the titles and abstracts of 7,101,375 papers were examined — . These were divided into 42,956,115 sentences.

5.2.1 *Definitions*

Enju defines phrase and token elements (See `http://www-tsujii.is.s.u-tokyo.ac.jp/enju/enju-manual/enju-output-spec.html`). A phrase may consist of either one token or one or more phrases. I define a *token* as being a chemical entity if (and only if) it is marked as such by the NER stage. I define a *phrase* as being a chemical entity if:

- all the tokens within it are part of a chemical entity
  e.g. *Hydrochloric acid* has two tokens, both of which are part of a chemical name; OR
- the semantic head (as defined by Enju) of the phrase is a chemical entity
  e.g. *The concentrated hydrochloric acid* has as its semantic head the noun phrase *concentrated hydrochloric acid*, which in turn has the semantic head *hydrochloric acid*; OR
- it is a coordination, at least one of the elements of which is a chemical entity.
  e.g. *Alcohol and tobacco* is a coordination, and because one of its elements (*alcohol*) is a chemical entity, it is considered for our purposes, even though the second element (*tobacco*) is not.

This definition is applied recursively, so if one of the elements in a coordination is a phrase whose semantic head is a chemical entity, we consider the coordination when looking for hyponyms.

A phrase is a candidate for being a hypernym ("descriptor") if:

- it is a noun phrase
  AND EITHER
- the first token within it is a determiner
  e.g. ***A*** *commonly used antibiotic* ; ***The*** *only pesticide*;
  OR

- it is a plural noun phrase.
  e.g. *Commonly used antibiotic**s***

5.2.2 *Extraction rules*

A phrase representing a chemical entity and one representing a potential hypernym are assigned as a pair if:

- there exists a token with the chemical and the descriptor as copular arguments 1 and 2 (in any order)
  e.g. *was* in *The most commonly used pesticide **was** DDT.* OR
- there exists a token with the chemical and the descriptor as appositive arguments 1 and 2 (in any order)
  e.g. the comma in *Amoxicillin**,** an antibiotic*

5.3 HYPERNYM RECOGNITION

The output is in standoff format, that is, it specifies the positions in the sentence at which XML tags should start and finish. This allows integration with information about chemical entities. As an example, the following information is derived from parsing the article title *Smokeless tobacco and tobacco-related nitrosamines.*[*]

The standoff output from Enju was reconstituted into XML, in combination with the named entity information, so that tokens which are part of a chemical entity are marked as such.

* Smokeless tobacco and tobacco-related nitrosamines. Cogliano *et al.*, Lancet Oncology, Dec 2004;5(12):708

```
,,
<enju ver="2.3.1">
0       51      sentence id="s0" parse_status="success" fom="7.7552"
0       50      cons id="c0" cat="NP" xcat="" head="c1" sem_head="c1" schema="
    empty_spec_head"
0       50      cons id="c1" cat="NX" xcat="COOD" head="c2" sem_head="c2" schema="
    coord_left"
0       17      cons id="c2" cat="NX" xcat="" head="c4" sem_head="c4" schema="mod_
    head"
0       9       cons id="c3" cat="ADJP" xcat="" head="t0" sem_head="t0"
0       9       tok id="t0" cat="ADJ" pos="JJ" base="smokeless" lexentry="[&amp;lt
    ;ADJP&amp;gt;]N_lxm" pred="adj_arg1" arg1="c4"
10      17      cons id="c4" cat="NX" xcat="" head="t1" sem_head="t1"
10      17      tok id="t1" cat="N" pos="NN" base="tobacco" lexentry="[D&amp;lt;N
    .3sg&amp;gt;]_lxm" pred="noun_arg0"
18      50      cons id="c5" cat="COOD" xcat="" head="c6" sem_head="c6" schema="
    coord_right"
18      21      cons id="c6" cat="CONJP" xcat="" head="t2" sem_head="t2"
18      21      tok id="t2" cat="CONJ" pos="CC" base="and" lexentry="[N&amp;lt;
    CONJP&amp;gt;N]" pred="coord_arg12" arg1="c2" arg2="c7"
22      50      cons id="c7" cat="NX" xcat="" head="c9" sem_head="c9" schema="mod_
    head"
22      37      cons id="c8" cat="ADJP" xcat="" head="t3" sem_head="t3"
22      37      tok id="t3" cat="ADJ" pos="JJ" base="tobacco-related" lexentry="[&
    amp;lt;ADJP&amp;gt;]N_lxm" pred="adj_arg1" arg1="c9"
38      50      cons id="c9" cat="NX" xcat="" head="t4" sem_head="t4"
38      50      tok id="t4" cat="N" pos="NNS" base="nitrosamine" lexentry="[D&amp;
    lt;N.3sg&amp;gt;]_lxm-plural_noun_rule" pred="noun_arg0"
</enju>
<chem_ent e="50" s="38">nitrosamines</chem_ent>
<plain>Smokeless tobacco and tobacco-related nitrosamines.</plain>
```

Listing 3: Output of Enju and named-entity recognizer. The first two columns of the output from Enju are the positions, in characters, around which the tag in the third column should be placed

The standoff in Listing 3 is converted into XML with a `<chem>` tag at each instance of a chemical name, and an `entity_list` attribute in each XML tag representing a token (`<tok>`) or phrase (`<cons>`) which is part of the entity. This facilitates searching for chemicals at a later stage.

,,

```
<sentence fom="7.7552" id="s0" parse_status="success">
  <cons cat="NP" head="c1" id="c0" schema="empty_spec_head" sem_head="c1" xcat="">
    <cons cat="NX" head="c2" id="c1" schema="coord_left" sem_head="c2" xcat="COOD"
         >
      <cons cat="NX" head="c4" id="c2" schema="mod_head" sem_head="c4" xcat="">
        <cons cat="ADJP" head="t0" id="c3" sem_head="t0" xcat="">
          <tok arg1="c4" base="smokeless" cat="ADJ" id="t0" lexentry="[&lt;ADJP&gt
               ;]N_lxm" pos="JJ" pred="adj_arg1">Smokeless</tok>
        </cons>
        <cons cat="NX" head="t1" id="c4" sem_head="t1" xcat="">
          <tok base="tobacco" cat="N" id="t1" lexentry="[D&lt;N.3sg&gt;]_lxm" pos=
               "NN" pred="noun_arg0">tobacco</tok>
        </cons>
      </cons>
      <cons cat="COOD" head="c6" id="c5" schema="coord_right" sem_head="c6" xcat="
           ">
        <cons cat="CONJP" head="t2" id="c6" sem_head="t2" xcat="">
          <tok arg1="c2" arg2="c7" base="and" cat="CONJ" id="t2" lexentry="[N&lt;
               CONJP&gt;N]" pos="CC" pred="coord_arg12">and</tok>
        </cons>
        <cons cat="NX" head="c9" id="c7" schema="mod_head" sem_head="c9" xcat="">
          <cons cat="ADJP" head="t3" id="c8" sem_head="t3" xcat="">
            <tok arg1="c9" base="tobacco-related" cat="ADJ" id="t3" lexentry="[&lt
                 ;ADJP&gt;]N_lxm" pos="JJ" pred="adj_arg1">tobacco-related</tok>
          </cons>
          <chem id="ent0"/>
          <cons cat="NX" entity_list="ent0" head="t4" id="c9" sem_head="t4" xcat="
               ">
            <tok base="nitrosamine" cat="N" entity_list="ent0" id="t4" lexentry="[
                 D&lt;N.3sg&gt;]_lxm-plural_noun_rule" pos="NNS" pred="noun_arg0">
                 nitrosamines</tok>
          </cons>
        </cons>
      </cons>
    </cons>
  </cons>
.</sentence>
```

Listing 4: Enju parse tree reconstructed into XML with named-entities labelled in a form that can be searched by XQuery. The added elements that indicate the presence of named entities are highlight in red.

The XML in Listing 4 is a representation of a parse tree, as shown in Figure 14.

Information in the XML attributes specifies that, for instance, token t2 ("and") is a predicate of type `coordination` with two arguments: c2 ("Smokeless tobacco") and c7 ("tobacco-related nitrosamines"). c7 has a `semantic-head` attribute that tells us that the head noun of the phrase is t4 ("nitrosamines"). t4 has a `base` attribute telling us that the lemma of the plural noun is "nitrosamine".

5.3.1 *XQuery*

Simple XQuery [27] queries representing copular and appositive relations are used to extract hypernym/hyponym pairs where the hyponym has been identified as a chemical entity. These implement the rules described in sections 5.2.1 and 5.2.2.

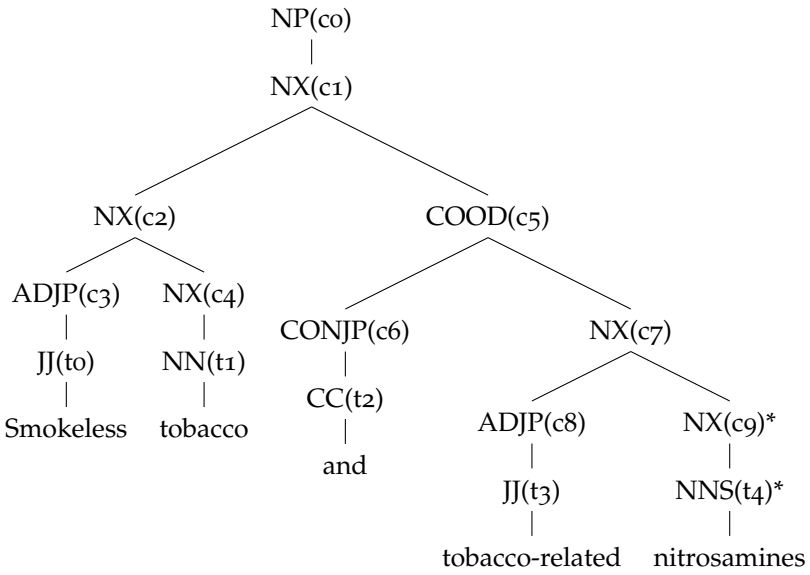


Figure 14: Parse tree for phrase "Smokeless tobacco and tobacco-related nitrosamines." Elements identified by NER as being within chemical entities are marked with an asterisk.

The use of XQuery to extract information from parse trees for Dutch text is detailed in Bouma and Kloosterman(2007)[20]; for this project a purpose-built XQuery library was composed.

The Java library Saxon* was used as an XQuery engine. It is capable of operating on flat files, thus not requiring the XML to be "shredded" into a relational database. XQuery samples are to be found in Appendix C.

5.3.2 *Normalization*

All descriptors longer than a single token have the potential for multiple readings. For example, consider the phrase:

a promising new antiviral drug

which is parsed as in Figure 15. Enju identifies the semantic head of the phrase as being the noun "drug". The candidate phrases are all those which include this token: "promising new antiviral drug", "new antiviral drug", "antiviral drug", and "drug". These options are evaluated in decreasing order of length, looking for a match to a ChEBI term. "promising new antiviral drug" and "new antiviral drug" do not match, but "antiviral drug" matches (exactly) to `CHEBI:36044`: "antiviral drug". This spares us from having to resolve the term to the over-general (albeit accurate) `CHEBI:23888`: "drug".

* `http://saxon.sourceforge.net/`

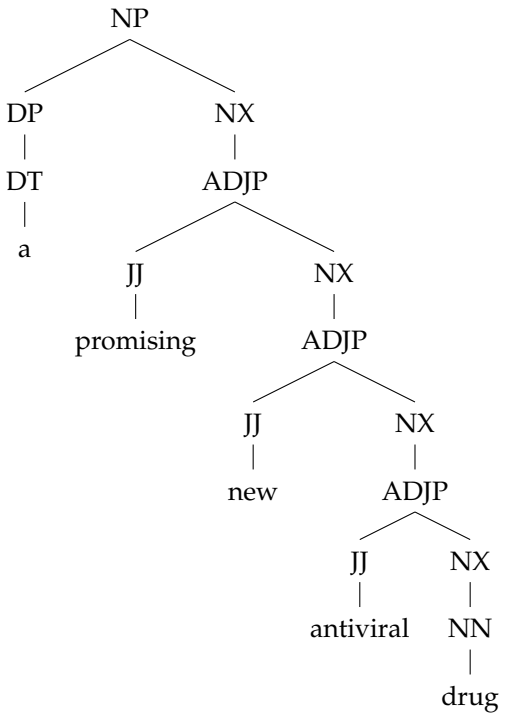


Figure 15: Parse tree for phrase "a promising new antiviral drug"

When resolving chemicals and descriptors to ChEBI terms, the string as found in the text is compared first to primary terms, then exact synonyms, then less exact synonyms, in that order. If no match is found, then the string is normalised (plural endings, letter case, and some non-alphanumeric characters are ignored), and the same comparisons repeated.

5.4 HUMAN EVALUATION

7.1 million PubMed abstracts were used in the generation of 74970 tuples of the form `{chemical,hypernym}` — where both the chemical and hypernym were nodes in the ChEBI ontology — and presented to ChEBI team members for manual verification. 954 randomly-selected tuples were manually annotated by two members of the ChEBI team.

The , who were very familiar with the ontology, were instructed to assess each tuple for correctness — that is, to state whether the ChEBI entities as identified by the software truly had a hypernymic relationship, regardless of whether the sentence which was the source of the tuple provided good evidence for this being the case.

5.4.1 *Evaluation Guidelines*

The definition of hypernym used was defined by:

- $(X\ is\_a\ Y) \wedge (Y\ is\_a\ Z) \rightarrow (X\ is\_a\ Z)$
- $(X\ has\_role\ Y) \rightarrow (X\ is\_a\ Y)$
- $(X\ has\_functional\_parent\ Y) \rightarrow (X\ is\_a\ Y)$
- $(X\ is\_conjugate\_base\_of\ Y) \rightarrow (X\ is\_a\ Y)$
- $(X\ is\_conjugate\_acid\_of\ Y) \rightarrow (X\ is\_a\ Y)$.

These last two items are due to the common practice in biological literature of using the names of conjugate acids and bases (say, "glutamate" and "glutamic acid") interchangably.

See Figure 17 for the guidelines as they were summarised for the annotators. Figure 16 shows the web annotation interface.

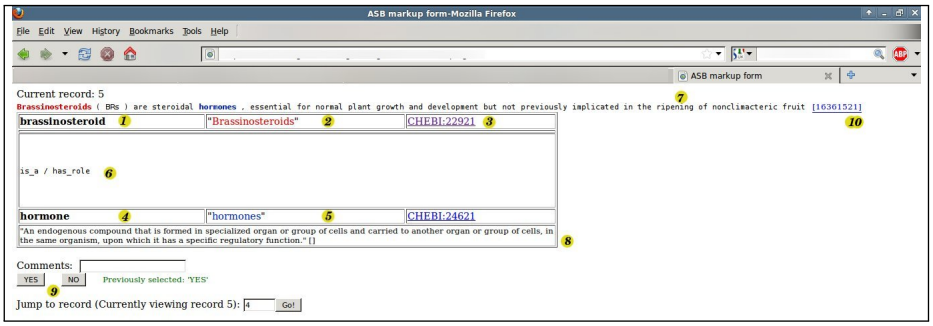


Figure 16: Interface for annotation. **1**: Canonical (ChEBI) name of chemical. **2**: Reference to chemical, as it appears in the sentence. **3**: ChEBI ID of the chemical. **4**: Canonical (ChEBI) name of hypernym. **5**: Reference to hypernym, as it appears in the sentence. **6**: Relationship type (for this exercise it does not vary). **7**: Sentence from which relationship was derived. **8**: Description (if present) in ChEBI of the relevant entity. **9**: Forced-choice annotation. **10**: PubMed ID and link to PubMed for full abstract.

5.4.1.1 *Conjugate acids and bases*

The source of the information from which we are extracting hyperymic relations is biological literature. It is a feature of biology that almost all reactions take place in an aqueous environment, and it has become conventional among biologists that the ionization state of a species that may exist in one of several states need not generally be specified. Thus we refer indifferently to *aspartate* or *aspartic acid*, for example.

ChEBI Annotation Guidelines 14 March 2011

1. For the purposes of this exercise, the relations "is_a" and "has_role" are not distinguished. In this document, I will use the → character for this relation. I will talk about the "Chemical" and the "Property". Pairs of chemicals and properties should be marked irrespective of whether any sentence shown provides adequate evidence for the relationship.

   Glutamic acid → Amino acid : **true**
   Glutamic acid → Neurotransmitter : **true**

2. If the sentence quoted does not provide evidence of a relationship that is nonetheless true, or states something to be true which is nonetheless false, add the comment "NS".

3. It is not relevant whether the Property is the best fit or the most specific description of the Chemical. If Chemical → X and X → Y, Chemical → Y.

   Quetiapine → Antipsychotic drug : **true**
   Quetiapine → Central nervous system drug : **true**
   Quetiapine → Drug : **true**
   Quetiapine → Molecule : **true**

4. Conjugate acids and Conjugate bases are considered to be identical.

   Glutamic acid → Amino acid : **true**
   Glutamate → Amino acid : **true**
   Zinc → Ion : **true**
   Zinc (2+) → Ion : **true**

5. If the Chemical is not a single compound but a class of compounds, the statement should be marked as true if-and-only-if the Property applies to members of the class in general.

6. The ChEBI description of the Chemical and the Property should be considered. For example, "Alcohol" is used in texts to refer to ethanol [CHEBI:16236], the class of organic hydroxyl compounds [CHEBI:33822], an -OH group, or any previously mentioned organic hydroxyl compound.
   Sulfapyridine → Pyridines [CHEBI:26421]: **true**
   Sulfapyridine → Pyridine [CHEBI:16227]: **false**
   Hydrogen → Gallium : **false** (the term "gas" can be mis-identified as the plural of "Ga").

7. Any ChEBI term → itself. For terms that are similar but not identical, bear in mind point above: Pyridine → Pyridines, but not vice versa.

The annotators should not make decisions on specific cases by comparing notes with each other or with the same third party.

Figure 17: The document that was provided as a summary of annotation procedures to the ChEBI curators.

5.4.1.2 *Systematic polysemy*

The "pyridine" example given in point 6 in the guidelines is taken from Corbett et al.(2008)[36], which discusses this variation of systematic polysemy within chemical names.

5.4.2 *Results*

Of the 954 tuples, annotator 1 (A1) assigned 747 (78%) as true; Annotator 2 (A2) assigned 712 (75%) as true.

5.4.2.1 *Assessing quality of annotation*

To assess whether the guidelines are coherent — that is, that they are clear, for a particular tuple, whether it should be annotated as "true" or "false" — we can compare the results of two human experts following the guidelines, to see how reproducible the results are. For coherent guidelines with good annotators, we would expect a high level of agreement — higher than chance alone. A low level of agreement would suggest either poor guidelines or poor annotators.

To measure this, we can use the Cohen's Kappa coefficient κ[102;96;24] which allows comparison between two annotators.

κ is calculated as

$$\frac{P_a - P_e}{1 - P_e}$$

where P_a, the observed agreement, is calculated as

$$\frac{\sum_i^n [A1_i = A2_i]}{n}$$

representing the proportion of the n tuples in which the two annotators obtain the same results; and P_e, expected chance agreement is calculated as

$$\frac{\sum_i^n [A1_i = \text{“}T\text{”}]}{n} \cdot \frac{\sum_i^n [A2_i = \text{“}T\text{”}]}{n} + \frac{\sum_i^n [A1_i = \text{“}F\text{”}]}{n} \cdot \frac{\sum_i^n [A2_i = \text{“}F\text{”}]}{n}$$

which would be reached by random annotation using the same distribution of classes as the human annotators.*

We can consider as a baseline, a "naive" annotation policy N that marks every tuple as the most frequent category: "true". This will agree with A1 78% of the time; the expected agreement by chance would therefore be

* Formulae assume that $A1_i$ is the judgement, either T or F, given by annotator A1 to tuple i

$(0.78 \times 1) + ((1 - 0.78) \times (1 - 1))$ or 0.78. κ is then calculated as $\frac{0.78-0.78}{1-0.78} = 0$. This illustrates the value of using Kappa over percentage agreement, which at 78% makes N look like a reasonable annotator.

There are various standards for interpretation of κ values. Krippendorff's scale rates $\kappa \geq 0.8$ as *reliable*, $0.67 \leq \kappa < 0.8$ as *marginally reliable*, and $\kappa < 0.67$ as *unreliable*. Landis & Koch, by contrast, rate $0.2 < \kappa \leq 0.4$ as *fair* correlation, $0.4 < \kappa \leq 0.6$ as *moderate*, $0.6 < \kappa \leq 0.8$ as *substantial*, and $\kappa > 0.8$ as *almost perfect*.[102]

In our case, A1 and A2 were in agreement for 903 (95%) tuples. Since the expected agreement by chance was $(0.78 \times 0.75) + ((1 - 0.78) \times (1 - 0.75))$, or 64%, we may calculate a κ of 0.85. This is a very high value[24], which reflects well on the use of annotators whose occupation is the curation of ChEBI and who are hence extremely well-trained in whether a statement meets the criteria for inclusion.

In those cases where the annotators disagreed, the author of this thesis acted as a third annotator in order to produce a consensus annotation, in which 727 (76%) of tuples were designated as correct.

The analysis was carried out at the level of individual tuples rather than distinct tuples since it allows us to consider features of the sentence from which the chemical and hypernym were extracted.

There were 808 distinct tuples (considering two tuples to be identical if their chemicals resolved to the same ChEBI term and their hypernyms resolved to the same ChEBI terms). 591 (73%) of these were correct according to the consensus annotation. This figure is slightly lower than that for non-distinct tuples since the tuples more commonly attested were more likely to be correct (see Section 5.4.3.4).

Of the 727 tuples designated as correct by the consensus annotation, 394 (54%) were also implied by the ChEBI ontology, and 333 (45%) were not. (None of the 227 tuples designated as incorrect was implied by ChEBI; if we examine distinct tuples, 312 (53%) of the 591 correct tuples were implied by ChEBI.)

5.4.3 *Features affecting accuracy*

5.4.3.1 *Length of chemical and hypernym names*

One noticeable feature of the data is the poor performance of resolution for very short chemical names. Of tuples involving one-character chemical names, all but one of the 16 annotated was incorrect. Of the 10 tuples with

two-character chemical names annotated, there was better performance with 8 (80%) correct. For three-character names, 9 out of 14 (64%) were correct.

Similarly, for hypernyms, all but one of the 11 tuples with single-character hypernyms (see Table 2) were marked incorrect. None of the 11 tuples with two-character hypernyms was marked correct, and only two (2.9%) of the 69 tuples with three-character hypernyms. Of these last, it should be noted that 55 of the 69 tuples had the hypernym “one" (or capitalised “One"), and four had “gas". These frequently-found and almost invariably incorrect hypernyms might be good candidates for blacklisting. For the two-character hypernyms, “mg" (milligrams, being misinterpreted as magnesium) and “mM” (millimolar, being misinterpreted as a pentasaccharide*) were frequently found.

Restricting the analysis to tuples that have both a chemical and a hypernym with length > 5 gives a precision of 87.2%, at the cost of decreasing the yield from 954 to 687 (hence reducing the recall by the same fraction).

5.4.3.2 *Syntactic relationship type*

It will be seen (Table 3) that relationships in which the chemical precedes the hypernym in the sentence (henceforth *forwards*) (“Chemical is a hypernym”, “Chemical, a hypernym...”) have better precision than those in which the hypernym precedes the chemical (henceforth *backwards*) (“The hypernym was Chemical”, “a hypernym, Chemical...”). Using just these "forward" tuples, we may obtain a precision of 77% for a yield of 703. Using these in combination with the term length restriction (length > 5 for both chemical and hypernym), we obtain a precision of 89% with a yield of 500.

5.4.3.3 *Systematic mis-resolution*

... there is a great variety of Equivocals. So the word Bill signifies both a Weapon, a Bird's Beak, and a written Scroll ... [112]

Some commonly-found hypernyms are frequently discovered to be false. The hypernym “one” is attested 51 times, all incorrect. This is due to the English number “one” being resolved as an abbreviation for the chemical name “(E)-4-oxonon-2-enal”. Another common word mis-resolved is “group”, resolved to the ChEBI term referring to a molecular substituent. The term “gas” appears four times in the set, each time mis-resolved as the plural of Ga, the abbreviation of “gallium”.

* α-D-Man-(1→3)-[α-D-Man-(1→6)]-β-D-Man-(1→4)-β-D-GlcNAc-(1→4)-D-GlcNAc; CHEBI:53458

Letter	Resolution	
A	hydrogen acceptor	CHEBI:13193
B	boron-11	CHEBI:52451
C	carbon-14 atom	CHEBI:36927
D	deuteride	CHEBI:29301
E	L-glutamic acid	CHEBI:16015
F	fluorine-19 atom	CHEBI:36940
G	guanine	CHEBI:16235
H	hydrogen atom	CHEBI:49637
I	iodine-129 atom	CHEBI:52636
K	potassium-39 atom	CHEBI:52632
L	L-leucine	CHEBI:15603
M	L-methionine	CHEBI:16643
N	nitrogen-14 atom	CHEBI:36938
O	oxygen-19 atom	CHEBI:36933
P	phosphorus-33 atom	CHEBI:37973
Q	L-glutamine	CHEBI:18050
R	hydrogen acceptor	CHEBI:13193
S	sulfur-38 atom	CHEBI:37985
T	thymine	CHEBI:17821
U	uranium atom	CHEBI:27214
V	vanadium atom	CHEBI:27698
X	xenon atom	CHEBI:49957
Y	yttrium-89 atom	CHEBI:52622
Z	benzyloxycarbonyl group	CHEBI:51097
a	hydrogen acceptor	CHEBI:13193
c	charm quark	CHEBI:36369
d	down quark	CHEBI:36367
e	electron	CHEBI:10545
h	helion	CHEBI:30220
k	potassium-39 atom	CHEBI:52632
n	neutron	CHEBI:30222
o	oxygen-19 atom	CHEBI:36933
p	proton	CHEBI:24636
s	strange quark	CHEBI:36368
w	tungsten-183	CHEBI:52462

Table 2: Single-letter terms resolved as chemicals attested in the tuple set

	forward	backward
appositive	77.4%	72.9%
copular	78.2%	67.8%

Table 3: Precisions of different syntactic relationship types for hypernyms

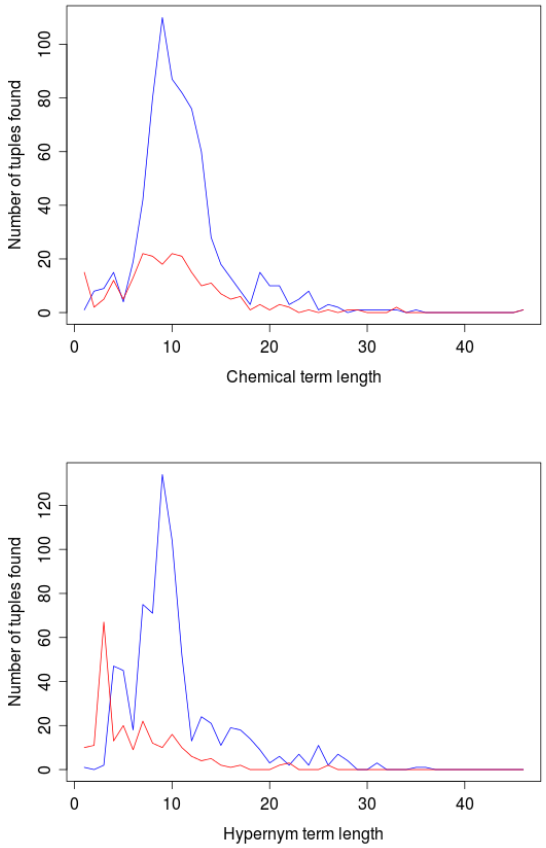


Figure 18: The effect of chemical and hypernym length on accuracy of tuples.Blue line indicates correct tuples; red line indicates incorrect tuples. It is seen that for very short chemical names, the number of false positives is large, exceeding the number of true positives. The effect is even more pronounced for hypernym names.

5.4.3.4 *Frequency*

One way that we can attempt to reduce non-systematic error is by adopting an error model as described by Downey et al.(2005)[41] and examining the frequency with which a tuple is attested. Frequencies were recorded across the whole set of extracted candidate sentences, not just the annotation set.

The most common tuples are given in Table 4, along with their frequencies.

For each of the tuples in the manually-annotated subset we take the total number of ocurrences in the superset. These are based on the ChEBI IDs of the chemical and hypernym, and ignore any variation in use of synonyms that may occur. We find that for tuples attested only once in the superset, 92 (60%) are incorrect – whereas for tuples attested more than five times, this figure falls to 9.7%.

Frequency	Tuple
359	nitrosyl is_a molecule
277	glutamate(2-) is_a neurotransmitter
256	rapamycin is_a inhibitor
239	trichostatin A is_a inhibitor
206	gamma-aminobutyric acid is_a neurotransmitter

Table 4: Most common tuples, along with their frequencies. All are correct by the definitions of section 5.4.1; commonly-attested tuples are much more likely to be correct than randomly-chosen tuples. The names shown for entities are the canonical ChEBI names.

5.4.3.5 *Scope of hypernym*

Tuple accuracy is affected by the level of generality (or "scope") of the hypernym in the tuple. Tuples containing more general terms are more likely to be correct – this is perhaps unsurprising since a term that covers a large proportion of chemicals (e.g. "drug") has a greater chance of being correct by chance than one that covers only a few chemicals (e.g. "antihypertensive drug").

With the caveat that the current coverage of ChEBI is not even across the domain of chemicals, we may estimate whether a given hypernym has a broad or narrow scope by the number of ChEBI nodes that it currently describes. The root node of ChEBI, "chemical entity", thus has the broadest scope, of 28881. By way of example, "sulfur molecular entity" has a scope of 2282, "organochlorine compound" has a scope of 667, and "steroid hormone" has a scope of 155. There are 20220 nodes that have a scope of 1 - since we consider every node as describing itself, these are leaf nodes.

Looking at the histograms of the scopes of the annotated hypernyms (Figures 19 and 20, we see that the hypernyms from incorrect tuples have a distribution skewed towards narrower scopes. This is what would be expected, since the broader the scope, the more chance there is that any given node is described by it – to give an extreme example, the broadest-scoped node "chemical entity" will describe correctly *any* chemical.

5.4.3.6 *Negations*

In designing the patterns to extract hypernymic relations, the issue of negations was considered. The patterns used will, given a sentence of the form *X is not a Y*, incorrectly extract the relation {*X* `is_a` *Y*}. Such sentences proved to be rare in the literature, and for this reason it was not considered necessary to attempt to detect and exclude negations. Of the 227 tuples annotated

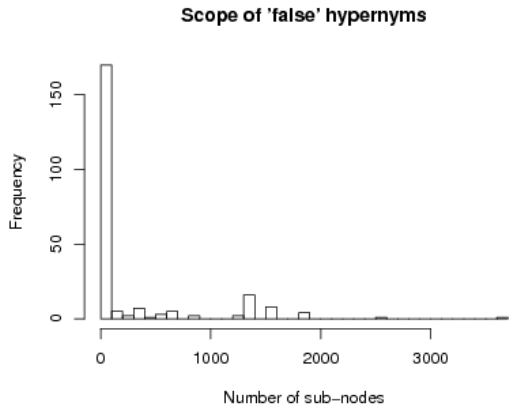


Figure 19: Scope of 'false' hypernyms. Incorrect tuples are heavily skewed towards very narrow scopes

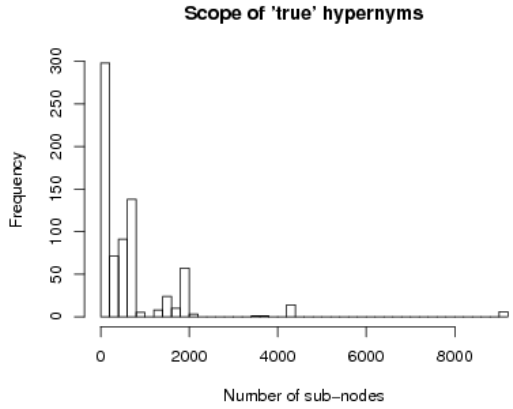


Figure 20: Scope of 'true' hypernyms. While correct tuples are skewed to narrow scopes, they are skewed to a much smaller extent than incorrect tuples

as being incorrect in the evaluation set, none appear (to the author of this thesis) to be incorrect due to undetected negations. Additionally, since there are multiple ways of writing in order to hedge or to negate statements, not all of which will involve the lexeme "not", the task of automatically detecting and excluding such sentences would not be trivial.

5.4.4 *Comparison with simpler lexicosyntactic patterns*

Our method of extracting the hypernyms is analagous to two of the so-called Hearst patterns from Kolářik et al.(2007)[60] (although not in Hearst(1992)[51]). We can examine which of our tuples might have been extracted using an implementation of this. The relevant patterns are:

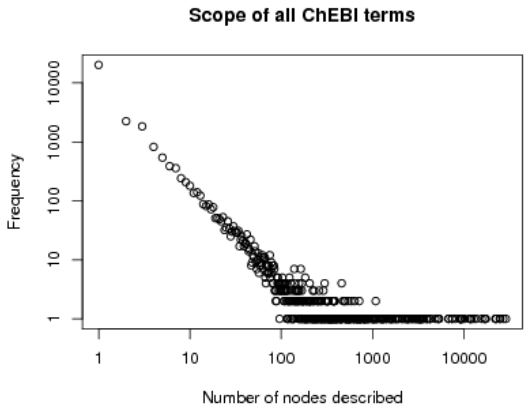


Figure 21: Scope of all ChEBI terms. In this log/log plot of frequency against scope, it is seen that the ChEBI ontology contains a large number of leaf nodes (scope=1) and nodes which describe only a small number of other nodes (narrow scope), and a handful of nodes which describe very many other nodes (broad scope)

NP_1 is (a | an) NP_0

NP_1, (a | an) NP_0 [60]

At a simple level we can investigate which of the "forward" tuples in our annotated set match the expression

[Chemical] (is | ,) (a | an) [hypernym].

Without a method of analysing the phrase structure of the sentences, the recall is poor: of the "forward" tuples that are known to be correct, 95 are matched by this expression and 607 are not.

Some of these undetected relations would be found by examining nested noun phrases as in Figure 15, for instance

> *"In addition , **catechol** **is** a major **metabolite** of carcinogenic benzene"*

— whereas other cases would not, for instance:

> *"**Tamoxifen** has not only proved to be a valuable treatment for estrogen receptor (ER) -positive breast cancer , but **is** also a pioneering **medicine** for chemoprevention in high-risk pre- and postmenopausal women"†*

would require a parser that can deduce that the verb "is" has as its predicate "Tamoxifen", and that can cope with the adverb "also" before the noun phrase containing "medicine". The large number of intervening phrases

* Oikawa S, Hirosawa I, Hirakawa K, Kawanishi S. Site specificity and mechanism of oxidative DNA damage induced by carcinogenic catechol. Carcinogenesis. 2001 Aug;22(8):1239-45.

† Park WC, Jordan VC. Selective estrogen receptor modulators (SERMS) and their roles in breast cancer prevention. Trends Mol Med. 2002 Feb;8(2):82-8.

would deter a system such as that used in Rindflesch and Fiszman(2003)[90] which filters by distance between verb and complement.

This, of course, ignores the extra recall that might be contributed by the other Hearst patterns, but Snow et al.(2005)[99] demonstrates (for general-domain text) that most of these have rather low recall compared to appositive and copular relationships.

5.5 AUTOMATIC PREDICTION OF TUPLE ACCURACY

Given we have shown there are several predictors of tuple quality, it would be useful to combine these into an overall predictor for tuple accuracy. For this purpose, we may use a standard logistic regresssion tool such as that included within the Weka machine-learning toolkit[114]. The use of this model allows us to ignore the independence assumption that methods such as Naïve Bayes would impose.

We use the features previously discussed:

- Scope of hypernym (numeric)
- Attestation frequency (numeric)
- Syntactic relation type — appositive *vs.* copular (category)
- Syntactic relation direction — forwards *vs.* backwards (category)
- The token which has the chemical and hypernym as its arguments. For appositive relations, this is a comma, dash, semicolon, or colon; for copular relations it is some form of the verb *to be*. (string)
- A boolean flag which is true if the hypernym is one of `{one gas mg mm group}`

For this exercise, it is important that we not include duplicate entries – otherwise any machine learning algorithm would be highly likely to simply learn the truth of individual tuples, which would be uninformative for the purpose of marking up novel tuples. To avoid this complication we use the reduced set of 808 distinct tuples; when there are two tuples that have the same Chemical ChEBI-ID and the same Hypernym ChEBI-ID one is selected arbitrarily.

Our naïve baseline (assume all results are true) gives a precision of 73% (for a recall of 100%; F-measure 85%).

With the default logistic regression settings of Weka* corresponding to a ridge parameter[26] of 10^{-8}, and iterating until convergence, running a 50-fold cross-validation, the precision rises to 87% with a recall of 96% (F-measure 91%). These default settings are set to maximise F, whereas for this project it is more useful to obtain a high precision. Since Weka yields a probability estimate for each data point, it is simple to set our own threshold. If, for instance, we set this at 0.95, we obtain a precision of 96% with a recall of 41%. Increasing the threshold further we can obtain a precision of 98% with a recall of 30%. Indeed, setting the threshold to 0.997, we can obtain a precision for this dataset of 100% with a recall of 18% (107 tuples retrieved).

5.5.1 *Trivial results*

There are two broad categories of trivial results - those where the hypernym is identical to the chemical, and those where the hypernym is over-general. The former case is not widespread - in the annotated set only two tuples were in this category. However, there were 24 cases of "molecule", which is not very informative given that we have restricted the semantic range of the hyponyms to chemical species.

5.5.2 *Examining the actual hypernyms tested*

In the list of most common hypernyms in Table 5, with the exception of "(E)-4-oxonon-2-enal", which was a false positive commented on above, all the terms here are reasonably general — some being so general as to verge on the uninformative. In particular "molecule" is not a useful piece of information: all the entities that we are interested in classifying are molecules (with the exception of chemical elements, which are a well-characterised set and for which we already have a fixed list). Other terms such as "drug" and "metabolite" are more useful, but ideally we would like to know what an entity is a drug *for* and what it is a metabolite *of*. In Chapter 6 I go on to examine ways of extracting more information about these over-general hypernyms by developing patterns to identify some common modifiers to these entities.

* `weka.classifiers.functions.Logistic -R 1.0E-8 -M -1`

Hypernym		Frequency
inhibitor	CHEBI:35222	25799
antagonist	CHEBI:48706	7668
drug	CHEBI:23888	7447
(E)-4-oxonon-2-enal	CHEBI:58972	7020
metabolite	CHEBI:25212	5795
agonist	CHEBI:48705	4369
protein polypeptide chain	CHEBI:16541	3691
molecule	CHEBI:25367	3230
antioxidant	CHEBI:22586	2139
group	CHEBI:24433	1970
antibiotic	CHEBI:22582	1621
atom	CHEBI:33250	1403
neurotransmitter	CHEBI:25512	1356
hormone	CHEBI:24621	1184
peptide	CHEBI:16670	1173
alpha-amino acid	CHEBI:33704	1143
lipid	CHEBI:18059	1039
alkaloid	CHEBI:22315	925
cofactor	CHEBI:23357	916
tyrosine kinase inhibitor	CHEBI:38637	893

Table 5: Most common hypernyms attested in the tuple set

5.6 SUMMARY

This project has proved useful in providing a partially-organised set of tuples for the ChEBI curators that can be used to make the project of assignments of edges in the ChEBI ontology both faster and more comprehensive.

Curators who are selecting edges to add to ChEBI should be presented with suggestions that are most likely to be correct (since incorrect suggestions waste their time). By the simple heuristic of looking first at the most commonly-attested hypernyms for a chemical species, curators can enrich their list of candidate edges for these. The other advantage of this approach is that those hypernyms most frequently mentioned in the biological literature are likely to be those that the biological science community is most interested in.

If a higher precision than this is required, then other heuristics can be used, such as the exclusion of very short chemical or hypernym names.

Since the system extracts chemical-hypernym pairs from individual sentences, it provides sources for its assertions, which fulfils an additional requirement of the ChEBI curatorial procedures. Without these sources, the ChEBI curators would be required to perform a separate literature-searching step before they could add a new edge to ChEBI.

The comparison of the results with ChEBI allows us to estimate the extent to which this technique could be used. From the annotated sample, 47% of those (distinct) tuples found to be correct are not yet included in ChEBI, . At an exceedingly rough estimate, then, this technique might be used to approximately double the coverage of ChEBI.

In the next chapter, I describe the extension of the techniques used above to detect other types of relationships than hypernymy.

6

IDENTIFICATION OF NON-HYPERNYMIC RELATIONS

In Chapter 5 I covered the extraction of hypernyms, corresponding to the `is_a` and `has_role` relationships in ChEBI. In this chapter, I use the same software framework with a different set of XQuery patterns and a broader set of entities to extract a varied set of relationships between a variety of different semantic types. I examine the results, identifying which relationships occur between which semantic types; this yields possible methodologies for selection of new edge types.

6.1 OTHER RELATIONS

One of the main directions in which the work in Chapter 5 could most obviously be extended was in the identification of new hypernyms, and specifically in the narrowing of the broader-scoped (and extremely common) hypernyms most frequently attested. Terms like "inhibitor" are almost always found in the literature with modifiers, either premodified ("*P* inhibitor"), or postmodified ("inhibitor of *P*") where *P* is typically a biological process or a protein (or other active molecule).

It would be useful for us to identify terms that are modified with respect to known biological processes or proteins. There are already collections and ontologies of proteins, protein families, and biological processes and activities[11]. We can use these to identify entities to which *P* may be resolvable. This may help us in deciding which modified terms are common enough in their own right to be useful additions to ChEBI. This approach may also help in the planned extension of ChEBI to include inter-ontology links.

Many of the ontologies in which we are interested are, like ChEBI, produced in Open Biological Ontologies format. This has the advantage that they are compatible in such a way that they can easily be combined[98;12]. In this chapter, for the task of identifying non-hypernymic relations, I will take advantage of the fact that the ontologies can be combined, by using three ontologies, Protein Ontology (PRO)[78], GO[11], and the acDO[81], in addition to ChEBI.

We thus want to find informative patterns that will connect ChEBI entities not just to other ChEBI entities, as our hypernymic patterns do, but to these other biological entities.

6.2 PATTERNS

Since the syntactic patterns used had to be hand-coded in XQuery[27], I decided to focus on more general patterns rather than developing very large numbers of very specific patterns that would only match a few sentences each.

6.2.1 *Subcategorisation of hypernyms*

In looking at the hypernyms most commonly attested from Chapter 5, we saw that a common pattern was

```
Chemical_X is_a Y [preposition] Entity_Z
```

— for example

"*Vanadate, an inhibitor of the Ca(2+) pump*"

or

"*catechol is a major metabolite of carcinogenic benzene*"

— and that the semantic category of Z is typically related to the value of Y. For example, "Treatment for" or "Risk factor for" tends to take an argument of type *disease*, whereas "Substrate of" tends to take an argument of type *protein*, or less freqently of *biological process*, and "Analogue of" takes an argument of type *chemical*.

An XQuery pattern was built to identify these triples. The same pattern as described in Chapter 5 was used to extract hypernyms, and X and Z were restricted to noun phrases that could be resolved to OBO entities (ChEBI, Human Disease Ontology (DO), GO, PRO); additionally triples where X resolved to anything other than a ChEBI entity were discarded. Y was restricted to noun phrases (as identified by Enju). Appendix A shows the semantic types of entities attaching to the most common values of Y and preposition.

The next task was to identify patterns to match cases where the hypernym is a compound noun, such as in

"*milnacipran , a serotonin-noradrenalin reuptake inhibitor*"

In these cases we cannot search (as we could in the previous case) for a preposition connecting the noun phrases. Enju does not perform well at parsing these phrases; it prefers a parse such as that shown in Figure 22 rather than the correct interpretation shown in Figure 23. Enju fails to recognize "serotonin-noradrenalin reuptake" as a noun phrase of its own, and as such it will not be selected for resolution. This appears to be a systematic idiosyncrasy.

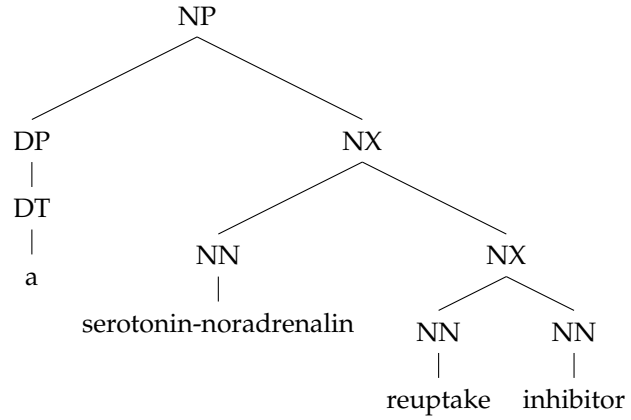


Figure 22: Enju's parse tree for phrase "a serotonin-noradrenalin reuptake inhibitor"

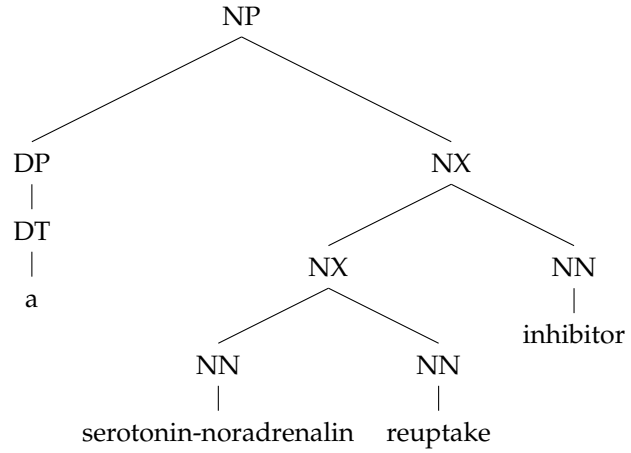


Figure 23: Desired parse tree for phrase "a serotonin-noradrenalin reuptake inhibitor"

The accomodation arrived on was to examine only single-token phrases, and for each hypernym of $n > 1$ tokens, to attempt to resolve the token sequence $t_{1,\ldots,(n-1)}$ to an OBO entity, considering t_n (if it is a noun*) to have the entity as a noun argument. Thus each phrase can only be split into an OBO entity covering all but the last token, followed by a single noun. Any phrase that could not be analysed in this way was rejected. These two-part

* Enju attribute `cat="N"`

noun phrases are referred to here as "NP2" (see Table 6). Some examples are shown in Figure 24.

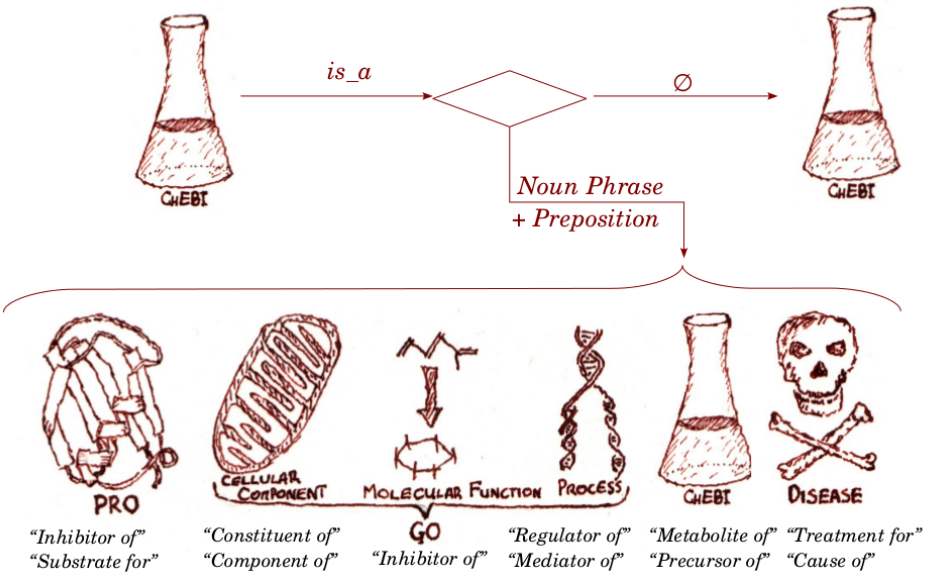


Figure 24: The treatment of hypernyms. If a hypernym is resolvable to a noun phrase, a preposition, and an entity within any of the ontologies which we are considering, then the relationship is considered valid. Some typical relationship types are shown for each ontology. This also applies to the two-part noun phrase pattern, where a hypernym is decomposable to an ontology entity followed by a single noun.

6.2.2 *Verb phrases*

`Entity_X [transitive verb] Entity_Y` with one of X or Y being a chemical (identified as such according to the NER system and resolving to a ChEBI term) — for example:

> "***Somatostatin*** *inhibits* ***secretion***"

or, in the other direction*,

> "*In contrast, the* ***hydrogenosome*** *of Trichomonas species metabolise* ***pyruvate*** *via a pyruvate : ferredoxin oxidoreductase*"

Some examples are shown in Figure 25.

* This latter relation type, with the chemical as the object, is indicated in this dissertation by a superimposed rightward arrow over the verb, *e.g.* for the example here, [Pyruvate $\overrightarrow{\text{VBS.metabolise}}$ hydrogenosome]

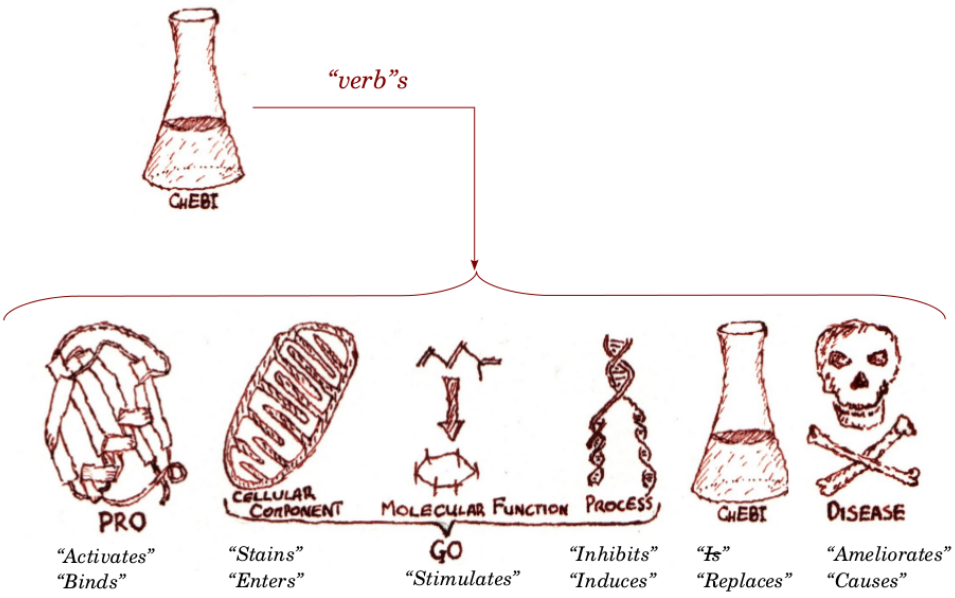


Figure 25: Verb phrases and semantic types. Some typical relationship types are shown for each ontology. Note that *Is* is crossed out, since relationships involving the copular verb are already included as hypernyms.

Type	Composition	Example
Hypernymic	HYPONYM	HYPONYM
Transitive verb (forwards)	VBS.verb	VBS.modify
Transitive verb (backwards)	$\overbrace{\text{VBS.verb}}$	$\overbrace{\text{VBS.contain}}$
Noun phrase	NP2.noun	NP2.analog
Noun phrase + preposition	PRP.noun␣phrase‖preposition	PRP.risk␣factor‖for

Table 6: Composition of abbreviations of terms.

6.3 HUMAN ANNOTATION

One of the questions we have to address before attempting annotation, is how to judge correctness. Many of the assertions are highly context-specific. For example:

> *The* ***organism*** *lacks* ***ergosterol*** *but contains distinct C28 and C29 delta7 24-alkylsterols*[*]

In this case "*the organism*" is anaphoric, referring to a previously mentioned organism — in this case, *Pneumocystis carinii*.It is not in general true that all organisms lack ergosterol; but it is true of at least one organism in at least one context. Besides anaphoric noun phrases, assertions may be specific to a subset of organisms, disease states, or experimental conditions.

[*] Kaneshiro ES, Wyder MA. C27 to C32 sterols found in *Pneumocystis*, an opportunistic pathogen of immunocompromised mammals. *Lipids*. 2000 Mar;35(3):317-24. PMID:10783009

In other cases, the meaning of a verb changes depending on the semantic type of the argument. For example, *inducing* a protein has a specific meaning (*i.e.* causing the gene which encodes the protein to be transcribed and translated), somewhat different to the more colloquial meaning of *inducing* applied to a biological process.

While the entities are resolved to nodes within ontologies, and the default assumption must be that annotators should be asked to construe the entities according to the descriptions in the said ontologies, the non-hypernymic relationship types, by and large, do not have standard definitions. For example:

> *Epidermal bioassay demonstrated that* ***benzylamine****, a membrane-permeable weak base, can mimick* ***hydrogen peroxide*** *(H2O2) to induce stomatal closure [. . .]**

In this case do we need to have a specific guideline determining how annotators should construe *mimick*, or is it possible to have an all-purpose protocol that can rely on the fact that our annotators are fluent in English?

The annotation was performed by the same annotators as in Chapter 4, through a similar web interface, modified to allow for the differences in the data structures. A small annotation was performed as a pilot study with no per-relation glosses, to investigate the Interannotator Agreement (IAA) when the annotators were instructed to use their intuition as to the meaning of the relations.

6.3.1 *Annotation Guidelines*

Guidelines given to annotators were as follows:

All annotation should be performed through the web interface, which has been kept as similar as possible to that used for the annotation of the `is_a`/`has_role` relations in 2011, on the assumption that the same annotators will be performing this annotation.

6.3.2 *Terms*

ENTITY An object (from an OBO ontology)

* Hydrogen peroxide-induced changes in intracellular pH of guard cells precede stomatal closure. Zhang X, Dong FC, Gao JF, Song CP. *Cell Res.* 2001 Mar;11(1):37-43.

ENTITY TYPE The category of an entity. For our purposes:

1. CHEBI chemical or chemical property
2. DO human disease
3. GO biological process
4. GO cellular component
5. GO molecular function
6. PRO protein

RELATION A word or phrase characterising the way that two entities interact.

RELATION TYPE The category of a relation. For our purposes:

1. Verb
2. Two-Part Noun Phrase.
3. Noun Phrase and Preposition

TRIPLE a set of three items consisting of: a CHEMICAL, a *Relation*, and an **Entity** (which may be of any of the listed Entity Types, including ChEBI).

6.3.3 *Principles of annotation*

6.3.3.1 *Forced Choice*

All triples should be marked True or False. There is a comment field in which to note any difficulties, but a judgement is still required on each triple.

6.3.3.2 *How to interpret entities*

Where a description is shown, an entity should be interpreted according to that description, which is taken from the relevant ontology. This is the case even if the term is being used in the sentence in a different sense, which may well render the triple False.

- e.g.: **precursor** can be resolved as **TGF-beta 1 isoform 1** (PR:000000397). Unless the triple is true interpreting it in this sense, it should be marked as False.

6.3.3.3 *How to interpret relation types*

*Verb*s are expected to be interpretable as *[*Chemical X `verbs` ***Entity Y****]* (or in some cases *[**Entity Y** `verbs` Chemical X]*).

- "The LP9 **cells** *produced* hyaluronic acid" — here "cell" is the entity and "hyaluronic acid" is the chemical
- "phenylacetate *induced* **transcription**" — here "transcription" is the entity and "phenylacetate" is the chemical

The form of the verb should *not* be taken into account - so "inhibit", "inhibits", "inhibited", "inhibiting" are all equivalent and if the triple would be True with any one of them then the triple should be marked as True.

*Two-Part Noun Phrase*s are expected to be interpretable as [Chemical X `is_a` **Entity Y** *noun*].

- "Sodium nitroprusside is a **nitric oxide** *donor*".

*Noun Phrase + Preposition*s are expected to be interpretable as [Chemical X `is_a` *noun_phrase preposition* **Entity Y**]. The noun phrase is most often just be one word long, but it can be longer. The prepositions are most often *of, for, in, with, than*.

- "Retinoids are a *Derivative of* **Vitamin A**".

6.3.3.4 *How to interpret relations*

Depending on the preliminary results, the most common relations may have glosses prepared, and should be interpreted accordingly. It is important to note that some relations have more than one sense depending on the Entity type.

In the cases without glosses, the relations should be considered in the light of the sentence quoted; any suggestions or observations that may be helpful in compiling glosses can be placed in the "comments" field.

As is the case for ChEBI's `has_role` relations, they should be marked as True if they can be true under any circumstances; they are nor required to be true under all circumstances.

6.3.3.5 *"Backwards" verbs*

Verbs listed as *backwards* are simply cases in which, instead of being understood as *[*Chemical X *verbs* ***Entity Y****]*, they should be understood as *[**Entity***

Y verbs CHEMICAL *X]*. In the interface the chemical and the verb have been exchanged so the sense should be more obvious.

6.3.3.6 *Consultation*

The annotators should not make decisions on specific cases by comparing notes with each other or with the same third party.

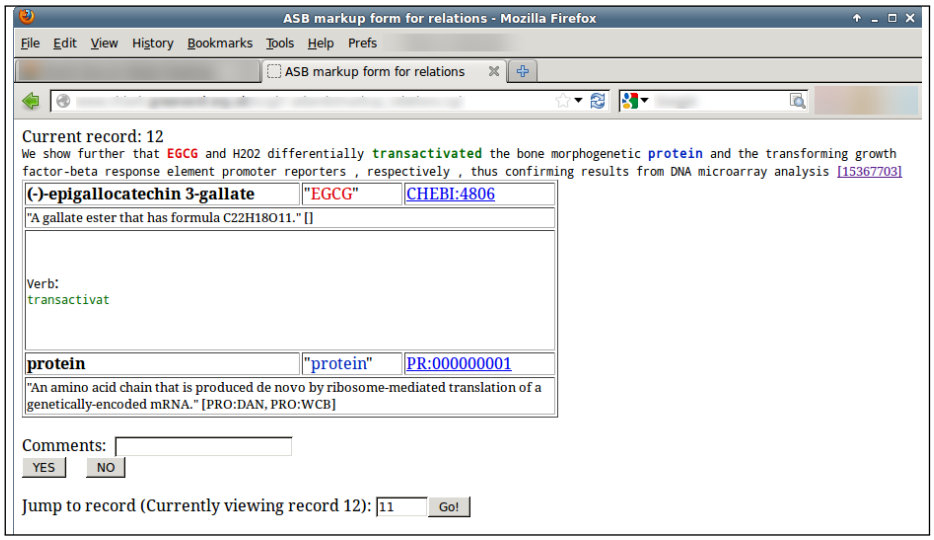


Figure 26: Annotation web interface for non-hypernymic relations. Note the presence of the verb in green where previously (Figure 16) there was the invariant relation "`is_a`".

6.3.4 *Results*

For the preliminary annotation, both annotators annotated 92 triples, each marking 31 incorrect and 61 correct.

This corresponds to a precision of 66.3%, and an IAA of 91.3%; $\kappa = 0.81$ This was high enough that the annotators were asked to proceed to the main body of annotation without requiring detailed per-relation guidelines, as mooted in Section 6.3.

	True	False
True	57	4
False	4	27

Table 7: Confusion matrix for first round of annotation. Rows are the results provided by Annotator 1; columns are those provided by Annotator 2 (though in this case the table happens to be symmetrical)

For the full annotation, 633 triples were annotated.

This corresponds to an average precision of 68.4%, and an IAA of 87%. Kappa is 0.71, which is much more disappointing than the preliminary annotation (and perhaps suggests that per-relation annotation would have been useful), but is still sufficient to have some confidence in the results.

The author acted as a third annotator (as in Chapter 4) to resolve disagreements and create a consensus annotation. The consensus result is 433 True, 200 False, corresponding to 68.4% precision.

The value for "forwards" verbs approaches the 77% achieved by Villaverde et al.(2009)[107] *after* filtering and combining triples (using a filter similar to that used by Kavalec and Svátek(2005)[58] for general domain relations). (Villaverde et al. do not give a result for precision for unfiltered triples).

If we restrict ourselves to triples attested in at least two separate papers, we find that precision rises to 83.8% although we have excluded 405 triples, leaving 228 in the annotated set. For "forwards" verbs the precision rises to 89%, with the yield falling to 63.

So far, we have deliberately been interested in extracting as broad a sample of relations as possible. A side-effect of this is that the data is somewhat sparse and rather noisy. We would like, if possible, to find unsupervised methods of reducing noise and sparsity in order that the data be more useful, if possible without limiting the scope of relations to match our preconceived notions of which events we think should be of interest to biologists.

	True	False
True	392	68
False	14	159

Table 8: Confusion matrix for annotation. Rows are the results provided by Annotator 1; columns are those provided by Annotator 2

74.2%	167	verb(forwards)	True
25.8%	58	verb(forwards)	False
69.8%	141	NP2	True
30.2%	61	NP2	False
67.7%	86	PRP	True
32.3%	41	PRP	False
49.4%	39	verb(backwards)	True
50.6%	40	verb(backwards)	False

Table 9: Precision by relation type

		Annotation	Filter
83.8%	191	True	Pass
16.2%	37	False	Pass
40.2%	242	True	Fail
59.8%	163	False	Fail

Table 10: Precision after filtering by number of attestations. To pass the filter, a triple had to be attested in two separate papers.

6.4 SEMANTIC PROFILES

We now have a set of relations, which we can think of as connecting a chemical to an ontological entity that is one of*:

- CHEBI
- DISEASE
- GO_BIOLOGICAL_PROCESS
- GO_CELLULAR_COMPONENT
- GO_MOLECULAR_FUNCTION
- PROTEIN

As we would expect, different relations do not connect to all of these identically, and most relations show clear affinities for one or for a few semantic types. The terms (abbreviated as described in Table 6) are seen in Appendix A.

* There are some relations that are shown as being from other types, notably species (shown in the appendices etc. as TAXON). This is due to the OBO files for some ontologies that were used containing definitions of entities from other compatible ontologies in order to be able to refer to them within their own definitions. Other (non-species) terms are grouped together as UNKNOWN, and mainly consist of data about ontologies, defining entities such as "node" and "property"

6.4.1 *Stemming*

In order to deal with the problem of data sparsity, such as items being excluded from analyses due to the infrequency of their occurrence, as well as for the sake of providing more coherent results, it is desirable to combine items when we can be reasonably sure that they represent the same information as each other. One way of doing this is to apply a stemming algorithm — here that of Porter(1980)[82] * — to combine relations which share a lexeme and vary only by tense, number, &c, such as *modify*, *modifies*, *modified*, *modifying*, which will all be transformed to *modify*. The implementation used also copes with some transatlantic spelling variants such as *recognize/recognise*† (though not *analog/analogue*, which was special-cased), with variations in capitalization, and with presence or absence of hyphens, often present in the 2-part noun phrases *X antagonist/X-antagonist*.

Some of these substitutions, such as PRP.vary‖of, which can be derived from any of (PRP.variability‖of, PRP.variant‖of, PRP.variation‖of, or PRP.variations‖of) are perhaps rather questionable as to whether distinct lexemes are being merged on the basis of sharing an etymological root. However, this problem is not specific to the use of stemming — there are similar issues of polysemy even without this process: compare the different uses of *group* noted in Chapter 5.

6.4.2 *Use of profiles for filtering*

One of the problems we have to deal with is false positive relations engendered by mis-parsing. For instance, we should identify *an inhibitor of pyruvate dehydrogenase* as PRP.inhibitor‖of the protein PR:000020907 "pyruvate dehydrogenase". If the boundaries of the noun phrases are mis-identified, we may end up identifying it as PRP.inhibitor‖of CHEBI:15361 "pyruvate". This is not only wrong, but semantically questionably coherent — it is not meaningful to talk of a chemical inhibiting another chemical. If we can identify the general semantic types of the arguments of relations, we may be able to filter out these aberrations.

A simple way to do this is to observe some obviously flawed relations and the prevalences at which they occur. Appendix A shows that the relation HYPONYM has 15% of its distinct arguments being proteins. Looking down

* as implemented by the Perl Text::English module `http://search.cpan.org/perldoc?Text::English`.
† This is an extension of Porter's algorithm by the author of the Perl module.

the table we see a general trend for such flawed relations to be more common for semantic types that make up a small percentage of that relation's arguments. Somewhat arbitrarily (but informed by the case of hypernyms and proteins) we may choose a figure such as 15% as a trade-off between generating clean data and the risk of losing some of the rarer relation types.

When we apply the filter to the annotated set, with a 15% threshold, we see a modest increase in precision (Table 11). The filter excludes 91 triples — just over half of which were annotated as false — retaining 542.

		Annotation	Filter
72%	390	True	Pass
28%	152	False	Pass
47.3%	43	True	Fail
52.7%	48	False	Fail

Table 11: Precision after filtering by semantic profile

6.5 POINTWISE MUTUAL INFORMATION

Something that we would like to consider is, given the data, can we identify relation types that are more-or-less equivalent? A useful empirical way of doing this is to use a measure such PMI.

If we take all relations between a given pair of a chemical and an entity (restricting our choice of entity to those of one semantic type at a time for the reasons described in 6.3), we can regard those relations as co-occurring (in the sense that they share a pair of arguments, not in the sense that they share a location in the literature; they may well not be mentioned in the same document).

So, if we see attested (Chemical C rel_1 Entity E) and (Chemical C rel_2 Entity E) and (Chemical C rel_3 Entity E) we can say rel_1 co-occurs with rel_2 ; rel_1 co-occurs with rel_3 ; rel_2 co-occurs with rel_3 (for the sake of simplicity we say that each of these co-occurences happens once per distinct pair of Chemical and Entity). If, for all attested pairs of relations, we calculate PMI(rel_1 , rel_2)

$$= \log \frac{p(rel_1, rel_2)}{p(rel_1)p(rel_2)}$$

or, calculating using counts of occurences

$$= \log \frac{\frac{C(rel_1, rel_2)}{n}}{\frac{C(rel_1)}{n} \cdot \frac{C(rel_2)}{n}}$$

$$= \log \frac{n \cdot C(rel_1, rel_2)}{C(rel_1)C(rel_2)}$$

then we will see the pairs of relations that are most likely to be synonyms with the highest values (limited to those that occur together at least ten times). These are summarised in Table 12. These might usefully form the basis of a data-driven synonym detector, perhaps augmented by an in-the-loop human or a semantic resource such as WordNet that can take into account the meanings of the terms (cf. for instance Mougin et al.(2006)[73]).

Casting a human eye over Table 12, we see it contains a variety of semantic and lexical relations.

- NP2.ligand & PRP.ligand||for: Although all the terms have been stemmed, they have not been combined between relation types.

PMI	Sem type	Relations		Together	Separately	
10.60	ChEBI	NP2.study	NP2.trial	11	53	22
10.08	GO_CC	VBS.releas	VBS.secret	12	52	35
9.967	GO_MF	NP2.antagonist	NP2.blocker	30	85	58
9.725	GO_CC	PRP.compon‖of	PRP.structur_compon‖of	15	162	18
9.625	GO_CC	VBS.gener	VBS.produc	12	21	119
9.600	GO_CC	NP2.inhibitor	PRP.inhibitor‖of	11	63	37
9.570	GO_CC	VBS.secret	VBS.synthes	11	35	68
9.551	GO_CC	VBS.produc	VBS.secret	19	119	35
9.447	GO_CC	PRP.compon‖of	PRP.lipid_compon‖of	11	162	16
9.320	GO_CC	NP2.constitu	PRP.compon‖of	17	27	162
9.235	PROTEIN	NP2.blocker	PRP.antagonist‖of	12	84	39
9.210	GO_CC	PRP.compon‖of	PRP.constitu‖of	28	162	48
9.158	GO_CC	NP2.compon	PRP.constitu‖of	11	66	48
9.129	PROTEIN	NP2.ligand	PRP.ligand‖for	19	90	62
9.124	GO_CC	VBS.releas	VBS.synthes	12	52	68
9.119	PROTEIN	NP2.ligand	PRP.ligand‖of	14	90	46
9.070	GO_CC	VBS.produc	VBS.utiliz	14	119	36
8.980	ChEBI	VBS.serv	VBS.suggest	13	51	83
8.969	GO_CC	VBS.accumul	VBS.take	16	101	52
8.940	PROTEIN	NP2.agonist	PRP.agonist‖of	20	216	31
8.910	ChEBI	NP2.regimen	NP2.therapy	11	38	99
8.899	DISEASE	PRP.factor‖for	PRP.risk_factor‖for	29	97	103
8.831	PROTEIN	NP2.study	NP2.trial	25	210	43
8.820	GO_CC	VBS.produc	VBS.releas	17	119	52
8.772	PROTEIN	NP2.activat	PRP.activat‖of	13	67	73
8.770	ChEBI	VBS.exert	VBS.serv	13	96	51
8.644	GO_CC	NP2.compon	PRP.compon‖of	26	66	162
8.593	GO_CC	VBS.produc	VBS.synthes	19	119	68
8.583	PROTEIN	NP2.antagonist	PRP.antagonist‖of	23	253	39
8.470	GO_BP	PRP.precursor‖for	PRP.substrat‖for	13	67	90
8.451	DISEASE	PRP.therapy‖for	PRP.treatment‖in	12	141	40
8.426	PROTEIN	NP2.agonist	PRP.agonist‖for	14	216	31
8.354	PROTEIN	NP2.activat	NP2.ligand	12	67	90
8.344	PROTEIN	NP2.antagonist	NP2.blocker	42	253	84
8.135	ChEBI	NP2.scaveng	PRP.scaveng‖of	25	209	70
8.113	ChEBI	NP2.therapy	NP2.treatment	19	99	114
8.081	PROTEIN	NP2.agonist	NP2.ligand	32	216	90
8.072	GO_BP	NP2.blocker	NP2.inhibitor	23	37	380
8.069	GO_CC	VBS.take	VBS.enter	18	52	212
8.068	ChEBI	VBS.exert	VBS.suggest	13	96	83
8.062	ChEBI	NP2.class	VBS.compris	12	104	71
8.001	ChEBI	PRP.analog‖of	PRP.analogu‖of	28	124	145

Table 12: Mutual information between relation types. The first column ("PMI") shows the pointwise mutual information between the two relations shown in the third and fourth columns as described in section 6.5. The second column indicates the semantic type of the second component in the relation (the first was always CHEBI). The fifth column indicates the number of co-occurrences; the sixth and seventh indicate the number of triples involving the relations in the second and third columns respectively.

- PRP.compon‖of & PRP.structur_compon‖of: These terms are substrings of each other; this presents an additional means of detection of such synonyms.
- NP2.therapy & NP2.treatment: These terms are reasonably close synonyms.
- PRP.analog‖of & PRP.analogu‖of: Transatlantic spelling variations. Note that "analogue" has been stemmed to "analogu".
- NP2.activat & NP2.ligand: While these terms are not synonyms, there may well be a biological relationship in that a chemical that activates a protein will almost certainly need to act as a ligand for the same protein in order to do so.

"If tin whistles are made of tin, what do they make fog horns out of?" - Lonnie Donegan

Such an approach may also help subcategorize NP2 arguments. It is well-established [38] that the components of noun phrases can have a variety of semantic relationships to each other. Recall that the design of the NP2 pattern was informed by the high prevalence of the pattern "Chemical `is_a` Entity inhibitor". There is indeed a class of these relations where "Chemical `is_a` Entity N" can be paraphrased as "Chemical `is_a` N of Entity" or "Chemical *N*s Entity". There is also a large class where "Chemical `is_a` Entity N" tends to be better paraphrased as "Chemical `is_a` N of type Entity" or even "Chemical `is_a` N and also an Entity". This last version points to a strategy to distinguish these — we can look at the PMI between the NP2 relation and the hypernymy relation. Table 13 shows the highest and lowest PMIs. Note that these are only shown where there is at least some co-occurrence, so most of the extremely low PMIs will have been excluded. In spite of this, we can see at the bottom of the table some examples of NP2s that take a ChEBI-type argument, but which do not imply hyponymy - for instance *X is a Y metabolite* does not imply *X is a Y**.

6.5.1 *Mis-resolution*

As discussed in Section 5.4.3.3, some terms are systematically mis-resolved. This (as we would expect) applies not just to chemicals but to entities from other ontologies. Some of the most commonly mis-resolved are: "precursor", "impact", "step", and "rôle". Very common words that are resolved (correctly or incorrectly) are shown in Table 15. We can exclude some of

* The very lowest score on Table 13, "acid", is there by virtue of the frequency of "amino acid", where "amino" is resolved to `CHEBI:46882 - amino group`.

PMI	Sem type	Relations		Together	Separately	
2.298	ChEBI	HYPONYM	NP2.agent	1495	33257	1503
2.267	ChEBI	HYPONYM	NP2.drug	984	33257	1011
2.207	ChEBI	HYPONYM	NP2.messeng	14	33257	15
2.154	ChEBI	HYPONYM	NP2.interact	18	33257	20
2.154	ChEBI	HYPONYM	NP2.choic	18	33257	20
2.087	ChEBI	HYPONYM	NP2.molecul	189	33257	220
2.067	ChEBI	HYPONYM	NP2.combin	50	33257	59
2.065	ChEBI	HYPONYM	NP2.family	11	33257	13
2.058	ChEBI	HYPONYM	NP2.regimen	32	33257	38
1.996	ChEBI	HYPONYM	NP2.medic	75	33257	93
1.996	ChEBI	HYPONYM	NP2.analges	25	33257	31
1.984	ChEBI	HYPONYM	NP2.present	12	33257	15
1.958	ChEBI	HYPONYM	NP2.option	11	33257	14
1.919	ChEBI	HYPONYM	NP2.medicin	13	33257	17
1.837	ChEBI	HYPONYM	NP2.ratio	13	33257	18
1.832	ChEBI	HYPONYM	NP2.substitut	18	33257	25
1.830	ChEBI	HYPONYM	NP2.composit	23	33257	32
1.809	ChEBI	HYPONYM	NP2.pigment	17	33257	24
1.806	ChEBI	HYPONYM	NP2.therapy	70	33257	99
			⋮			
0.364	ChEBI	HYPONYM	NP2.inhibitor	51	33257	196
0.360	ChEBI	HYPONYM	NP2.analogu	103	33257	397
0.290	ChEBI	HYPONYM	NP2.blocker	22	33257	89
0.274	ChEBI	HYPONYM	NP2.complex	11	33257	45
0.205	ChEBI	HYPONYM	NP2.metabolit	86	33257	369
-0.174	ChEBI	HYPONYM	NP2.antagonist	57	33257	318
-0.315	ChEBI	HYPONYM	NP2.alkaloid	13	33257	80
-1.115	ChEBI	HYPONYM	NP2.donor	21	33257	225
-1.547	ChEBI	HYPONYM	NP2.acid	11	33257	159

Table 13: Mutual information between hypernyms and NP2 relations. The highest and lowest PMI scores are shown

these most freqent terms automatically; however the gain in precision is not great. Excluding the 500 most frequent lemmas, we see precision rise to 70.4% (416 correct, 175 incorrect) by excluding 42 of the triples in our annotated set.

Term	Entity	Canonical term
precursor	PR:000000397	TGF-beta 1 isoform 1
mice	PR:000005054	caspase-14
step	PR:000013460	tyrosine-protein phosphatase non-receptor type 5
impact	PR:000009019	protein IMPACT
role	CHEBI:50906	role

Table 14: Some frequently mis-resolved terms

Frequency Rank	Number of triples	Term (lower case)	Entity	Canonical term
2	10	be	CHEBI:30501	beryllium atom
4	1	of	CHEBI:30241	fluorosyl group
5	240	a	PR:000001069	transient receptor potential cation channel TRPV4 isoform 1
5	309	a	PR:000004372	agouti-signaling protein
9	2	to	PR:000016214	tryptophan 2,3-dioxygenase
11	24	i	CHEBI:17596	inosine
11	220	i	CHEBI:52636	iodine-129 atom
15	4	he	CHEBI:37004	helium-8 atom
15	2	he	DOID:13413	hepatic encephalopathy
20	30	this	PR:000024065	sulfur carrier protein ThiS
22	1	at	CHEBI:2666	amitriptyline
22	6	at	PR:000016450	transmembrane protease serine 11D
25	3	his	CHEBI:15971	L-histidine
25	1	his	PR:000008437	histidine ammonia-lyase
28	1	not	PR:000011336	homeobox protein notochord
28	3	not	PR:000011409	nuclear receptor subfamily 4 group A member 2
31	6	or	CHEBI:29287	gold atom
31	2	or	PR:000001497	opioid receptor
34	1	go	DOID:3086	gingival overgrowth
36	4	can	CHEBI:53439	calcineurin
39	2	if	CHEBI:5864	ifosfamide
67	1	him	CHEBI:16069	1H-imidazole
92	455	no	CHEBI:16480	nitric oxide
92	26	no	CHEBI:33396	nobelium
92	111	no	CHEBI:35801	nitroso group

Frequency Rank	Number of triples	Term (lower case)	Entity	Canonical term
93	108	man	PR:000007572	formin-like protein 2
98	285	many	PR:000023162	mannose permease IIC component
103	2261	one	CHEBI:58972	(E)-4-oxonon-2-enal
127	4	in	CHEBI:30430	indium atom
128	2	as	CHEBI:27563	arsenic atom
128	1	as	DOID:7147	ankylosing spondylitis
128	1	as	PR:000004386	argininosuccinate synthase
128	4	as	PR:000022165	arylsulfatase
129	99	last	PR:000015650	single-stranded DNA-binding protein 3
150	1	put	CHEBI:17148	putrescine
154	1	on	PR:000015475	SPARC
159	20	great	PR:000001667	relaxin receptor 2
160	239	same	CHEBI:15414	S-adenosyl-L-methionine
161	3	big	PR:000017395	WD repeat-containing protein 5
162	2446	group	CHEBI:24433	group
173	10	hand	PR:000027804	basic helix-loop-helix transcription factor HAND
195	3	so	CHEBI:45822	sulfur monoxide
210	4	mr	CHEBI:25354	mineralocorticoid
210	3	mr	PR:000001400	mannose receptor
210	48	mr	PR:000011407	mineralocorticoid receptor
213	1	hold	PR:000022919	DNA polymerase III subunit psi
220	140	large	PR:000009667	glycosyltransferase-like protein LARGE1
221	465	all	CHEBI:37690	allose
226	430	water	CHEBI:15377	water

Frequency Rank	Number of triples	Term (lower case)	Entity	Canonical term

Table 15: Common words in general English and how they are resolved, based on the top 250 lemmas in American English from the Corpus of Contemporary American English (`http://corpus.byu.edu/coca/`). Note that not all of these are necessarily incorrect: *water* and *group* are frequently correctly detected. Multiple resolutions for the same word are due to differences in capitalization.

In this chapter I have described the development and use of lexicosyntactic patterns to capture, with a reasonable level of precision, non-hypernymic relations, without restricting their scope to a predetermined set of interaction types. I also have investigated some means by which these relations can be characterized. In the next chapter I investigate to what extent these data can be used to infer similarities between these relationships.

7

IDENTIFICATION OF RELATIONS BY INFERENCE

In previous chapters I have concentrated on extracting and resolving explicitly-expressed relationships from biological literature. In this chapter, I consider the possibility of deriving implicit relationships from the dataset that has been so derived. These may be well-known to the scientific community, but not explicitly stated in the literature in terms that the existing software will extract, or they may be less well-characterised, providing hypotheses for future biological research.

There has been rather more interest in using properties to cluster, classify, and assign new properties to biological entities, than in characterising the properties[115]. This is particularly the case in the domain of chemistry and pharmacology, with work such as Gurulingappa et al.(2009)[47] assigning drug functions, and within the gene product domain covered by GO - see for example Couto et al.(2007)[37] measuring similarities between proteins based on their annotations with GO terms.*

In contrast, this chapter describes methods for identifying relationships between properties, which may be of the `is_a` or `has_role` type, and as such may be identified directly with a node in an ontology — but which also may involve a different interaction type discovered from the literature, as discussed in Chapter 6. We are interested in discovering links that will be useful to us in constructing and refining ontologies, and also perhaps in identifying biologically interesting hypotheses.

7.1 IMPLICIT KNOWLEDGE

In Chapter 6, we identified a set of triples of the form `{chemical, predicate, entity}`, where "entity" is a node mentioned in one of the ontologies: ChEBI, PRO, DO, and GO, and "chemical" is a node from ChEBI. "predicate" can be any of the types extracted as described in Chapter 6, enumerated in Table 6.

We can combine the predicate and the entity to form a single *property*, allowing us to make two-entry tuples such as: `{CHEBI:15365, VBS.reduce, GO:0042311}` ⇒ `{Aspirin, "reduce GO:0042311"}` (*Aspirin reduces vasodilation*),

* Although within the same paper, the authors do propose an interesting semantic similarity measure for GO terms based not on their allocation to different proteins, but on graph topology within the ontology; they then use this in measuring similarities between proteins.

or

`{CHEBI:15365, HYPONYM, CHEBI:35222}` ⇒ `{Aspirin, "is_a CHEBI:35222"}`
(*Aspirin is an inhibitor*).

Given a large number of these tuples, we can describe each property with a vector where each dimension is a chemical species, representing how many times in the corpus the property is applied to each chemical for all attested chemical species with a ChEBI ID (see examples in Table 16). Alternatively, we can use a binary vector where each dimension is either 1 or 0, depending on whether the property is attested for that chemical or not (see Table 17). In the studies in this chapter, we will use binary vectors unless otherwise specified.

	Aspirin	Morphine
"is_a inhibitor"	57	3
"reduces vasodilation"	1	0

Table 16: Two properties are shown here, each described by a vector of length two. "is_a inhibitor" is characterised by [57,3] and "reduces vasodilation" by [1,0].

	Aspirin	Morphine
"is_a inhibitor"	1	1
"reduces vasodilation"	1	0

Table 17: In this table, binary vectors are shown, in which only the presence or absence of attestations is relevant, not the frequency of occurrences within the corpus. Here "is_a inhibitor" is characterised by [1,1] and "reduces vasodilation" by [1,0].

7.1.1 *Pointwise Mutual Information*

In Section 6.5 we used Pointwise Mutual Information (PMI) to identify relations that shared chemical-entity pairs and were possibly conflatable. We can use a very similar methodology to look for properties (that is, relation-entity pairs) that share chemicals and are possibly semantically related. If we regard all pairs of properties that share a chemical as co-occurring (again, co-occurring once per distinct ChEBI term that they share), we can identify possible implications.

Note that this method does not readily identify terms that are nested as in Figure 27 (Section 7.2). Instead it tends towards identifying synonyms where properties ϕ_1 and ϕ_2 pertain to identical sets of chemical species.

The pairs of properties with the highest mutual information values are shown in Table 18.

PMI	Property 1			Property 2			Together	Separately	
14.944	NP2.analog	CHEBI:17761	ceramide	PRP.analog‖of	CHEBI:17761	ceramide	5	5	5
14.944	HYPONYM	CHEBI:63920	artemisinin derivative	NP2.deriv	CHEBI:223316	(+)-artemisinin	5	5	5
14.944	HYPONYM	CHEBI:61073	oxygen radical	NP2.radic	CHEBI:25805	oxygen atom	5	5	5
14.944	HYPONYM	CHEBI:38461	carbamate insecticide	NP2.insecticid	CHEBI:13941	carbamate	5	5	5
14.944	HYPONYM	CHEBI:37335	MRI contrast agent	NP2.agent	CHEBI:37335	MRI contrast agent	5	5	5
14.944	HYPONYM	CHEBI:36980	pyridine nucleotide	NP2.nucleotid	CHEBI:16227	pyridine	5	5	5
14.944	HYPONYM	CHEBI:34956	ruthenium red	NP2.red	CHEBI:30682	ruthenium atom	5	5	5
14.944	HYPONYM	CHEBI:24921	isoquinoline alkaloid	NP2.alkaloid	CHEBI:16092	isoquinoline	5	5	5
14.944	HYPONYM	CHEBI:24654	hydroxy fatty acid	NP2.fatty_acid	CHEBI:43176	hydroxy group	5	5	5
14.681	NP2.analogu	CHEBI:17489	3′,5′-cyclic AMP	PRP.analog‖of	CHEBI:17489	3′,5′-cyclic AMP	5	6	5
14.681	NP2.agent	CHEBI:35498	diuretic	NP2.drug	CHEBI:35498	diuretic	5	5	6
14.681	HYPONYM	CHEBI:64571	NMDA receptor agonist	NP2.agonist	GO:0004972	NMDA sel. glu. receptor activity*	5	6	5
14.681	HYPONYM	CHEBI:55347	vitamin K antagonist	NP2.antagonist	CHEBI:28384	vitamin K	6	6	6
14.681	HYPONYM	CHEBI:51373	GABA agonist	NP2.agonist	CHEBI:16865	gamma-aminobutyric acid	6	6	6
14.681	HYPONYM	CHEBI:50445	adenosine deaminase inhibitor	NP2.inhibitor	PR:000003707	adenosine deaminase	6	6	6
14.681	HYPONYM	CHEBI:4356	desferrioxamine B	NP2.deferoxamin	CHEBI:38157	iron chelator	6	6	6
14.681	HYPONYM	CHEBI:35441	antiinfective drug	NP2.agent	CHEBI:35441	antiinfective drug	5	6	5
14.681	HYPONYM	CHEBI:35137	hemoprotein	NP2.protein	CHEBI:30413	heme	5	6	5
14.681	HYPONYM	CHEBI:33320	actinoid atom	NP2.agent	CHEBI:33320	actinoid atom	6	6	6
14.681	HYPONYM	CHEBI:26440	pyrimidine nucleoside	NP2.nucleosid	CHEBI:16898	pyrimidine	6	6	6
14.459	NP2.composit	CHEBI:35366	fatty acid	NP2.fraction	CHEBI:18059	lipid	5	7	5
14.459	NP2.agent	CHEBI:50864	insulin-sensitizing drug	NP2.drug	CHEBI:50864	insulin-sensitizing drug	5	7	5
14.459	HYPONYM	CHEBI:8364	prazosin	NP2.prazosin	CHEBI:48706	antagonist	5	7	5
14.459	HYPONYM	CHEBI:63962	Hsp90 inhibitor	NP2.inhibitor	PR:000025350	HSPC protein	6	7	6
14.459	HYPONYM	CHEBI:62868	hepatoprotective agent	NP2.agent	CHEBI:62868	hepatoprotective agent	6	7	6
14.459	HYPONYM	CHEBI:61950	microtubule-stabilising agent	NP2.agent	CHEBI:61950	microtubule-stabilising agent	7	7	7
14.459	HYPONYM	CHEBI:59282	kappa-opioid receptor agonist	NP2.agonist	PR:000001590	kappa-type opioid receptor	5	7	5

Table 18: Highest mutual information pairs of properties. Note that the relation is commutative ($PMI(x,y) \equiv PMI(y,x)$); the ordering of Property 1 and Property 2 is arbitrary (alphabetization-based).
*_N-methyl-D-aspartate selective glutamate receptor activity_ was abbreviated to fit in the available width

Looking at Table 18, one of the first things we notice is the tendency of the PMI method to turn out redundant terms where, due to the parsing method, a common string will be interpreted by different patterns. So, taking the second row in Table 18 as an example, the sentence *X is an artemisinin derivative* will be parsed by the HYP pattern as the triple [X, `is_a`, artemisinin␣derivative] and by the NP2 pattern as the triple [X, NP2.derivative, artimisinin]. This may be problematic if we take this to be two pieces of evidence rather than one piece interpreted in two ways. A possibility for reducing this tendency could be only to accept NP2 patterns where the entire NP will not resolve to an entity, or at any rate will not resolve to the same entity as the subphrase. In many other cases either the entities are closely similar and the relationship types are identical (*e.g.* [`is_a` oxygen radical] ⇔ [`is_a` oxygen atom]) or the entities are identical (*e.g.* [`is_a` antiinfective drug] ⇔ [`is_a` antiinfective drug `agent`]). In spite of this tendency, the terms are generally not incorrect and may be useful for identifying closely-related properties.

In some respects this is effectively carrying out the type of lexical analysis alluded to in Bada and Hunter(2007)[13], Burgun and Bodenreider(2005)[22] and Ogren et al.(2004)[80], where links between ontologies are made on the basis of lexical patterns. So we see in Table 18 that [HYPONYM NMDA␣receptor␣agonist] has been found equivalent to [NP2.agonist N-methyl-D-aspartate␣selective␣glutamate␣receptor␣activity], for instance.

To examine what other relations are found besides the type of re-analysis of noun phrases as both a hypernymic and an NP2 pattern, we can exclude NP2 relations from our results. The new pairs with the highest Mutual Information are as shown in Table 19. We see that these are still very much akin to synonyms, some providing potentially useful definitions, such as [`is_a` fluorochrome ⇔ VBS.stains DNA] and [`is_a` mineralocorticoid ⇔ $\overleftarrow{\text{VBS.bind}}$ mineralocorticoid receptor]

PMI	Property 1			Property 2			Together	Separately	
14.418	PRP.form‖of	CHEBI:27470	folic acid	PRP.form‖of	CHEBI:37445	folate	5	6	6
14.266	HYPONYM	CHEBI:25354	mineralocorticoid	$\overleftarrow{\text{VBS.bind}}$	PR:000011407	mineralocorticoid receptor	5	8	5
14.096	PRP.excitatory_neurotransmitt‖in	CHEBI:35470	central nervous system drug	PRP.excitatory_transmitt‖in	CHEBI:35470	central nervous system drug	5	9	5
14.096	HYPONYM	CHEBI:50913	fixative	VBS.fix	GO:0005623	cell	5	9	5
14.003	HYPONYM	CHEBI:48432	angiotensin II	HYPONYM	CHEBI:61016	angiotensin receptor antagonist	5	8	6
14.003	HYPONYM	CHEBI:48354	polar solvent	VBS.contain	CHEBI:62947	ammonium acetate	5	8	6
13.944	PRP.donor_prefer‖for	GO:0019509	L-Met salvage from MeThAd*	VBS.replac	CHEBI:25017	leucine	5	5	10
13.833	PRP.ligand‖for	PR:000013058	PPAR gamma*	PRP.ligand‖of	PR:000013058	PPAR gamma*	5	6	9
13.833	PRP.class‖of	CHEBI:22582	antibiotic	PRP.group‖of	CHEBI:22582	antibiotic	5	9	6
13.807	VBS.advanc	GO:0009058	biosynthetic process	VBS.antagon	GO:0009058	biosynthetic process	5	5	11
13.807	PRP.nutry‖for	GO:0040007	growth	PRP.nutry‖for	OBI:0100026	organism	5	11	5
13.781	PRP.constitu‖of	GO:0005886	plasma membrane	PRP.constitu‖of	GO:0016020	membrane	5	7	8
13.681	PRP.product‖from	CHEBI:17234	glucose	PRP.end_product‖of	GO:0006113	fermentation	5	5	12
13.681	HYPONYM	CHEBI:33364	platinum	HYPONYM	CHEBI:33749	platinum molecular entity	5	6	10
13.681	HYPONYM	CHEBI:25681	omega-3 fatty acid	HYPONYM	CHEBI:36825	azido group	6	9	8
13.681	HYPONYM	CHEBI:25029	leukotriene	PRP.group‖of	CHEBI:25029	leukotriene	5	10	6
13.681	HYPONYM	CHEBI:24043	flavones	PRP.phenol‖in	CHEBI:33290	food	5	10	6
13.611	PRP.metabolit‖of	CHEBI:16480	nitric oxide	PRP.product‖of	CHEBI:16480	nitric oxide	5	7	9
13.588	VBS.replac	CHEBI:16811	methionine	VBS.replac	CHEBI:32535	histidine residue	5	8	8
13.588	$\overleftarrow{\text{VBS.contain}}$	CHEBI:24396	glycopeptide	$\overleftarrow{\text{VBS.contain}}$	CHEBI:27026	toxin	5	8	8
13.566	PRP.product‖from	CHEBI:17234	glucose	PRP.product‖of	GO:0006113	fermentation	5	5	13
13.566	PRP.attractant‖of	CHEBI:33709	amino acid	VBS.replac	CHEBI:27570	histidine	5	5	13
13.544	$\overleftarrow{\text{VBS.include}}$	CHEBI:16670	peptide	VBS.contain	CHEBI:16670	peptide	5	6	11
13.529	HYPONYM	CHEBI:11814	3-H-3-MG CoA*	PRP.inhibitor‖of	PR:000008636	3-H-3-MG CoA* reductase	6	8	10
13.511	VBS.replac	CHEBI:27897	tryptophan	VBS.replac	CHEBI:29016	arginine	8	9	12
13.459	VBS.replac	CHEBI:26271	proline	VBS.replac	CHEBI:28044	phenylalanine	5	10	7
13.459	VBS.replac	CHEBI:25017	leucine	VBS.replac	CHEBI:28044	phenylalanine	5	10	7
13.459	HYPONYM	CHEBI:51217	fluorochrome	VBS.stain	GO:0005574	DNA	5	14	5
13.459	HYPONYM	CHEBI:17822	serine	VBS.substitut	CHEBI:29016	arginine	5	14	5

Table 19: Highest mutual information pairs of properties, excluding NP2 relations. Note that the relation is commutative ($PMI(x,y) \equiv PMI(y,x)$); the ordering of Property 1 and Property 2 is arbitrary (alphabetization-based).
*_L-methionine salvage from methylthioadenosine_, _peroxisome proliferator-activated receptor_ and _3-hydroxy-3-methylglutaryl-CoA_ were abbreviated to fit in the available width

7.1.2 *Cosine similarity*

Cosine similarity[92] is a measure of the cosine of the angle between two vectors, and is a popularly-used metric to identify most similar pairs of vectors where the vectors have not necessarily been normalised to be the same length. It is calculated as:

$$\text{cosine similarity} = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} = \frac{\sum_{i=1}^{n} \vec{x}_i \vec{y}_i}{\sqrt{\sum_{i=1}^{n} (\vec{x}_i)^2} \sqrt{\sum_{i=1}^{n} (\vec{y}_i)^2}}$$

When the value is 1 the vectors $\vec{x}$ and $\vec{y}$ are identical in direction. When the value is zero the two vectors are orthogonal and have no shared dimensions. Table 20 contains the most similar pairs of properties (excluding NP2s) by this measure. These appear to be somewhat less general than those generated by a Mutual Information measure, and contain a number of useful pairs such as [is_a triazoles ⇔ PRP.agent‖for fungal_infectious_disease].

CS	Property 1			Property 2			Together	Separately	
0.833	PRP.form‖of	CHEBI:27470	folic_acid	PRP.form‖of	CHEBI:37445	folate	5	6	6
0.800	VBS.bind	PR:000011407	mineralocorticoid_receptor	VBS.convert	GO:0003845	11-β-HSD_NADP_activity*	4	5	5
0.791	HYPONYM	CHEBI:25354	mineralocorticoid	VBS.bind	PR:000011407	mineralocorticoid_receptor	5	8	5
0.770	VBS.replac	CHEBI:27897	tryptophan	VBS.replac	CHEBI:29016	arginine	8	9	12
0.767	HYPONYM	CHEBI:35664	HMG-CoA_reductase_inhibitor*	PRP.inhibitor‖of	PR:000008636	3H3M-HMG-CoA_reductase*	10	17	10
0.745	PRP.excitatory_NT‖in*	CHEBI:35470	central_nervous_system_drug	PRP.excitatory_transmitt‖in	CHEBI:35470	central_nervous_system_drug	5	9	5
0.745	HYPONYM	CHEBI:50913	fixative	VBS.fix	GO:0005623	cell	5	9	5
0.731	HYPONYM	CHEBI:53756	HIV-1_RT_inhibitor*	HYPONYM	CHEBI:59897	reverse_transcriptase_inhibitor	10	11	17
0.730	VBS.contain	PR:000021936	signal_peptide	VBS.have	PR:000014678	alpha-1-antitrypsin	4	6	5
0.730	VBS.caus	CHEBI:58972	E_-4-oxonon-2-enal	VBS.form	GO:0030315	T-tubule	4	5	6
0.730	PRP.reduct‖of	CHEBI:33709	amino_acid	VBS.replac	CHEBI:16449	alanine	4	5	6
0.730	PRP.micronutry‖for	OBI:0100026	organism	PRP.statu‖of	CHEBI:46662	mineral	4	5	6
0.730	HYPONYM	CHEBI:22660	aspartic_acid	VBS.substitut	PR:000001343	CD69_molecule	4	6	5
0.728	HYPONYM	CHEBI:33712	N-terminal_amino-acid_residue	HYPONYM	CHEBI:33715	N-terminal-α-aar*	11	12	19
0.722	HYPONYM	CHEBI:48432	angiotensin_II	HYPONYM	CHEBI:61016	angiotensin_receptor_antagonist	5	8	6
0.722	HYPONYM	CHEBI:48354	polar_solvent	VBS.contain	CHEBI:62947	ammonium_acetate	5	8	6
0.707	PRP.donor_prefer‖for	GO:0019509	L-met_salvage_from_MeThAd*	VBS.replac	CHEBI:25017	leucine	5	5	10
0.707	HYPONYM	CHEBI:25681	omega-3_fatty_acid	HYPONYM	CHEBI:36825	azido_group	6	9	8
0.680	PRP.ligand‖for	PR:000013058	PPARG*	PRP.ligand‖of	PR:000013058	PPARG*	5	6	9
0.680	PRP.class‖of	CHEBI:22582	antibiotic	PRP.group‖of	CHEBI:22582	antibiotic	5	9	6
0.676	VBS.replac	CHEBI:28044	phenylalanine	VBS.substitut	CHEBI:15356	cysteine	4	7	5
0.676	VBS.increas	CHEBI:48705	agonist	VBS.medy	GO:0046903	secretion	4	7	5
0.676	VBS.consum	NCBITaxon:10116	Rattus_norvegicus	VBS.drank	PR:000005054	caspase-14	4	7	5
0.676	PRP.product‖from	CHEBI:17234	glucose	PRP.product‖from	GO:0006113	fermentation	4	5	7
0.676	PRP.ferment_product‖from	CHEBI:17234	glucose	PRP.product‖from	CHEBI:17234	glucose	4	7	5
0.676	PRP.compon‖of	PR:000015198	LACaaT-2*	PRP.constitu‖of	PR:000015198	LACaaT-2*	4	7	5
0.676	PRP.attractant‖of	CHEBI:33709	amino_acid	VBS.replac	CHEBI:28044	phenylalanine	4	5	7
0.676	HYPONYM	CHEBI:35727	triazoles	PRP.agent‖for	DOID:1564	fungal_infectious_disease	4	7	5
0.676	HYPONYM	CHEBI:26051	phosphoamino_acid	VBS.hydrolyz	GO:0016791	phosphatase_activity	4	5	7
0.676	HYPONYM	CHEBI:22632	arsenic_molecular_entity	PRP.form‖of	CHEBI:27563	arsenic_atom	4	7	5

Table 20: Highest cosine similarity scores for pairs of properties, excluding NP2 relations. Note that the relation is commutative ($CS(x,y) \equiv CS(y,x)$); the ordering of Property 1 and Property 2 is arbitrary (alphabetization-based).
**11-β-hydroxysteroid_dehydrogenase_NAD_P_activity*, *3-hydroxy-3-methylglutaryl-coenzyme_A_reductase*, *HIV-1_reverse_transcriptase_inhibitor*, *L-methionine_salvage_from_methylthioadenosine*, *N-terminal_alpha-amino-acid_residue*, *PRP.excitatory_neurotransmitt‖in*, *hydroxymethylglutaryl-CoA_reductase_inhibitor*, *low_affinity_cationic_amino_acid_transporter_2*, *peroxisome_proliferator-activated_receptor_gamma* were abbreviated to fit in the available width

7.2 HYPOTHESIS TESTING

For the set of properties Φ we can test the proposition that, for any pair of properties $(\phi_1, \phi_2) \in \Phi$, one implies the other. That is, for any chemical $c \in C$:

$$has\_property(c, \phi_1) \implies has\_property(c, \phi_2)$$

As an example, we would expect that

$$has\_property(c, [is\_a\ psychoactive\ drug]) \implies has\_property(c, [affects\ central\ nervous\ system])$$

We shall denote ϕ_1 as the (putative) *inner term* and ϕ_2 as the (putative) *outer term*, according to the visualisation of their arrangement as in the Venn diagram Figure 27.

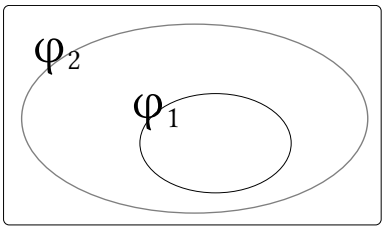


Figure 27: Venn diagram demonstrating innner and outer terms. ϕ_1 is the *inner term* and ϕ_2 is the *outer term*. Note that the only thing that makes them so is the fact that ϕ_1 does not apply to any chemicals to which ϕ_2 does not apply; there is not necessarily any taxonomic relationship between ϕ_1 and ϕ_2, and both may be contingent properties of the chemicals they describe.

Consider the vectors as described above, that is: $\mathbf{v}_{\phi,c}$ is the number of times that property ϕ is attested for chemical c.

A naïve approach might be to say, for each pair of properties $(\phi_1, \phi_2) \in \Phi$, we consider it to be supported if for all nonzero elements in $\mathbf{v}_{\phi_1}$, the corresponding element in $\mathbf{v}_{\phi_2}$ is also nonzero, and otherwise we consider it to be unsupported. This has two potential problems, however, both arising from sparsity of the corpus. Firstly we will reject pairs of properties where there exists even one chemical species that is described by the first and not described by the second. Secondly (and perhaps less problematic), we will accept pairs of properties where the inner happens to describe a very small number of chemicals and the outer is a hugely general term — such a pair will likely be uninformative.

Using all the detected relations (unfiltered by semantic profile; see Section 6.4) and arbitrarily restricting both inner and outer terms to those that both describe at least four different chemicals, we have 3709 pairs of properties that meet these criteria. As we would expect, these are mainly those

where the outer term is very general. Almost a third of these (1051) have as the outer term "is_a inhibitor"; the next most common outer terms are "is_a drug"(410) "is_a molecule"(289) "is_a metabolite"(189) "is_a ligand"(152) and "inhibits biosynthetic process"(127)*. — see Table 21 for more. The inner terms, (also as we would expect) have a skew to very specific terms that describe a very small number of chemicals. 725 of them describe only three chemical species and just a handful describe more than ten — see Table 22 for a more complete list.

Our first approach may be characterised as "no false negatives". In practice, we are aware that there are only too many false negatives.† One possible solution to this is to adopt a probabilistic model to help us answer the question: *to what extent is absence of evidence evidence of absence?*.

Let us consider a simple example, with a hypothetical inner property (ϕ_{Inner}), a hypothetical outer relation(ϕ_{Outer}), and four chemicals *A*, *B*, *C*, and *D*, with occurrences as shown in Table 23; shown graphically in Figure 28. We want to know how plausible it is that *D* is in the Outer group, even if it has not been attested in the literature as such.

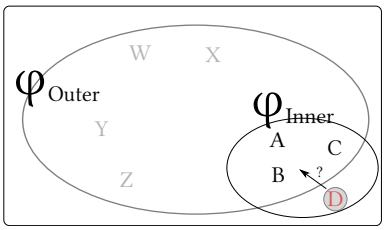


Figure 28: Venn diagram for comparison to Figure 27 demonstrating innner and outer terms for the data shown in Table 23. ϕ_{Inner} is the putative inner term and ϕ_{Outer} is the putative outer term. Chemicals *A*, *B*, and *C* are in both sets, supporting the hypothesis. Chemical *D* is only found in ϕ_{Inner}, opposing the hypothesis. (Chemicals *W*, *X*, *Y*, and *Z* are shown to emphasize that the hypothesis that ϕ_{Outer} is actually the *inner* term is not very plausible.) The arrow indicates our query: how plausible do we find it that *D* is *really* in both ϕ_{Inner} and ϕ_{Outer}, but has just not been observed in ϕ_{Inner} yet?

If we assign a fixed probability (say 50%) that anything that is not attested is actually untrue, then one way of looking at the results is that we have a 50% chance that all chemicals that are within ϕ_{Inner} are actually within ϕ_{Outer} (and if there were two chemicals in the same situation as *D*, the chance would be 25%). There's an intuitive problem with this, in that we have the same result if we have *A*, *B*, and *C* supporting the hypothesis, and *D* oppos-

* The GO term GO:0009058: Biosynthetic Process is attested in our data set by resolution of its synonyms (in order of prevalence) "formation", "synthesis", and "biosynthesis"

† To demonstrate that many such exist, however, we need only look at the fact that Chapter 5's hypernym-seeking did not rediscover all the hypernymic relationships implied by ChEBI.

Relationship type	Entity	Name	Chems	Freq
VBS.inhibit	GO:0005488	binding	327	29
VBS.induce	GO:0009058	biosynthetic process	288	31
HYPONYM	CHEBI:33250	atom	216	46
HYPONYM	CHEBI:48705	agonist	483	67
HYPONYM	CHEBI:33709	amino acid	186	84
VBS.inhibit	GO:0040007	growth	563	95
HYPONYM	CHEBI:48706	antagonist	543	98
VBS.inhibit	GO:0009058	biosynthetic process	517	127
HYPONYM	CHEBI:52214	ligand	569	152
HYPONYM	CHEBI:25212	metabolite	1099	189
HYPONYM	CHEBI:25367	molecule	762	289
HYPONYM	CHEBI:23888	drug	1172	410
HYPONYM	CHEBI:35222	inhibitor	1957	1051

Table 21: Most frequent outer terms, with their frequencies and the number of chemicals they describe (in column labelled "Chems")

N° species described	Frequency
4	725
5	386
6	176
7	131
8	75
9	46
10	34
11	24
12	15
13	14
14	11
15	8
16	1
17	5
18	2
19	3
20	2
25	1
29	1

Table 22: Numbers of chemicals described by inner terms

Chemical	ϕ_{Inner}	ϕ_{Outer}
A	✓	✓
B	✓	✓
C	✓	✓
D	✓	

Table 23

ing it, as if there were *A*, *B*, *C*, *E*, *F*, *G*, and an arbitrarily large number of other chemicals within both ϕ_{Inner} and ϕ_{Outer} and *D* opposing.

Instead, we can look at the chance that any given chemical with property ϕ_{Inner} also has property ϕ_{Outer}. In our example, the arithmetic (assuming that we take a simple arithmetic mean with no weighting) is $\frac{1+1+1+0.5}{4} = 0.875$. This method has the advantage that a large number of chemicals supporting the hypothesis will be given more weight than a small number, and that as the outcome of this calculation tends towards 1, we become more certain that the two sets are arranged as in Figure 27.

7.2.1 *Apriori*

A widely-used technique for generating association rules on a similar basis to this is the use of the Apriori[3;4] algorithm. This takes a set of instances, which in our case correspond to chemicals, and generates for them a candidate set of inferred rules, of the form ($Body_1$, $Body_2$, $\cdots$ $Body_n \Rightarrow$ Head). For reasons of simplicity, we will restrict ourself to n=1; that is, rules of the form (Body $\Rightarrow$ Head).

For each potential rule, Apriori emits figures for *Support* and *Confidence*. The rule X $\Rightarrow$ Y holds with confidence c if c% of chemicals that have property X also have property Y. The rule X $\Rightarrow$ Y has support s if s% of chemicals have both properties X and Y.[3]

Table 24 shows some of the rules that were discovered using this approach. The minimum support for a rule was set low, at 0.05% and the results were ordered by confidence. There are 2888 rules with 100% confidence. Table 25 shows some of the 1665 rules with 100% confidence that do not involve NP2 relations. The samples of rules shown in Tables 24 and 25 are selected arbitrarily* as a representative sample of the rules found.

* With GNU `shuf`, which shuffles the lines of a file

% Conf	Inner Property			Outer Property			Together	Separately	
100	VBS.contribut	CHEBI:52214	ligand	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	NP2.specy	CHEBI:16247	phospholipid	HYPONYM	CHEBI:18059	lipid	6	6	151
100	VBS.elevat	CHEBI:18243	dopamine	HYPONYM	CHEBI:23888	drug	6	6	1222
100	VBS.inhibit	GO:0030168	platelet_activation	VBS.inhibit	GO:0005488	binding	5	5	336
100	NP2.antagonist	CHEBI:50113	androgen	HYPONYM	CHEBI:35497	androgen_antagonist	8	8	18
100	NP2.vasodil	PR:000004372	agouti-signaling_protein	VBS.inhibit	GO:0009058	biosynthetic_process	5	5	519
100	HYPONYM	CHEBI:28918	R_-adrenaline	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	NP2.agent	CHEBI:38867	anaesthetic	HYPONYM	CHEBI:38867	anaesthetic	33	33	65
100	VBS.reduc	GO:0002118	aggressive_behavior	HYPONYM	CHEBI:23888	drug	5	5	1222
100	VBS.coordin	CHEBI:27363	zinc_atom	NP2.acid	CHEBI:46882	amino_group	6	6	147
100	HYPONYM	CHEBI:28971	ampicillin	HYPONYM	CHEBI:35222	inhibitor	6	6	2113
100	NP2.vasodil	PR:000004372	agouti-signaling_protein	HYPONYM	CHEBI:35620	vasodilator_agent	5	5	99
100	$\overbrace{\text{VBS.inhibit}}$	CHEBI:33232	application	HYPONYM	CHEBI:35222	inhibitor	7	7	2113
100	NP2.antagonist	PR:000001251	angiotensin_II_receptor_1A	NP2.blocker	PR:000001251	angiotensin_II_receptor_1A	8	8	8
100	NP2.carbohydr	GO:0051235	maintenance_of_location	HYPONYM	CHEBI:16646	carbohydrate	5	5	71
100	NP2.agent	CHEBI:35480	analgesic	HYPONYM	CHEBI:35480	analgesic	17	17	78
100	PRP.class‖of	CHEBI:18059	lipid	HYPONYM	CHEBI:18059	lipid	15	15	151
100	NP2.agent	CHEBI:61950	microtubule-stabilising_agent	HYPONYM	CHEBI:61950	microtubule-stabilising_agent	7	7	7
100	VBS.reduc	DOID:576	proteinuria	HYPONYM	CHEBI:23888	drug	8	8	1222
100	NP2.insecticid	CHEBI:13941	carbamate	HYPONYM	CHEBI:38461	carbamate_insecticide	5	5	5
100	HYPONYM	CHEBI:46631	clonidine	HYPONYM	CHEBI:23888	drug	5	5	1222
100	VBS.reduc	CHEBI:27698	vanadium_atom	VBS.stimul	GO:0009058	biosynthetic_process	5	5	192
100	PRP.metabolit‖of	CHEBI:27897	tryptophan	HYPONYM	CHEBI:25212	metabolite	7	7	1150
100	PRP.donor_prefer‖for	GO:0019509	L-met_salvage_from_MeThAd*	HYPONYM	CHEBI:33709	amino_acid	5	5	200
100	NP2.agent	CHEBI:37335	MRI_contrast_agent	HYPONYM	CHEBI:37335	MRI_contrast_agent	5	5	5
100	NP2.agent	CHEBI:38068	antimalarial	HYPONYM	CHEBI:35222	inhibitor	9	9	2113
100	HYPONYM	CHEBI:32460	cysteine_residue	HYPONYM	CHEBI:33709	amino_acid	5	5	200
100	NP2.activity	CHEBI:33282	antibacterial_agent	HYPONYM	CHEBI:22582	antibiotic	6	6	268
100	HYPONYM	CHEBI:32460	cysteine_residue	NP2.acid	CHEBI:46882	amino_group	5	5	147
100	NP2.protocol	CHEBI:35705	immunosuppressive_agent	HYPONYM	CHEBI:35222	inhibitor	5	5	2113

Table 24: Apriori results with confidence = 100%. These mainly fit into the category of the type of synonyms (discussed in Section 7.1.1) where a phrase can be parsed either as a single hypernym or broken into two parts as a NP2 pattern

**L-methionine_salvage_from_methylthioadenosine* was abbreviated to fit in the available width

% Conf	Inner Property			Outer Property			Together	Separately	
100	VBS.destabil	CHEBI:33699	messenger_RNA	HYPONYM	CHEBI:35222	inhibitor	6	6	2113
100	VBS.augment	GO:0016310	phosphorylation	HYPONYM	CHEBI:35222	inhibitor	8	8	2113
100	HYPONYM	CHEBI:15552	prostaglandin_I2	HYPONYM	CHEBI:23888	drug	8	8	1222
100	VBS.counteract	GO:0009058	biosynthetic_process	HYPONYM	CHEBI:35222	inhibitor	7	7	2113
100	VBS.prevent	DOID:4	disease	HYPONYM	CHEBI:23888	drug	12	12	1222
100	PRP.scaveng‖of	CHEBI:25941	peroxynitrite	HYPONYM	CHEBI:22586	antioxidant	6	6	260
100	PRP.substrat‖for	GO:0008610	lipid_biosynthetic_process	HYPONYM	CHEBI:25212	metabolite	5	5	1150
100	VBS.regul	GO:0045292	mRNA_cis_splicing_via_spliceosome	HYPONYM	CHEBI:35222	inhibitor	7	7	2113
100	PRP.target‖of	CHEBI:26523	reactive_oxygen_species	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	PRP.regul‖of	GO:0016042	lipid_catabolic_process	HYPONYM	CHEBI:52214	ligand	5	5	590
100	PRP.second_messeng‖in	GO:0007165	signal_transduction	HYPONYM	CHEBI:25367	molecule	7	7	659
100	VBS.replac	SO:0000418	signal_peptide	HYPONYM	CHEBI:61159	conjugated_linoleic_acid	5	5	359
100	HYPONYM	CHEBI:28001	vancomycin	$\overbrace{\text{VBS.rece}}$	CHEBI:24433	group	5	5	472
100	VBS.affect	CHEBI:38876	pyraclofos	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	VBS.replac	CHEBI:27594	carbon_atom	HYPONYM	CHEBI:33250	atom	5	5	228
100	$\overbrace{\text{VBS.self-administ}}$	NCBITaxon:10116	Rattus_norvegicus	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	$\overbrace{\text{VBS.bind}}$	PR:000011407	mineralocorticoid_receptor	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	PRP.product‖of	GO:0019369	arachidonic_acid_metabolic_process	HYPONYM	CHEBI:25212	metabolite	5	5	1150
100	VBS.reduc	GO:0042311	vasodilation	HYPONYM	CHEBI:35222	inhibitor	6	6	2113
100	PRP.inhibitor‖of	GO:0003824	catalytic_activity	HYPONYM	CHEBI:35222	inhibitor	15	15	2113
100	VBS.reduc	GO:0045730	respiratory_burst	VBS.inhibit	GO:0040007	growth	6	6	604
100	VBS.activat	GO:0006915	apoptotic_process	HYPONYM	CHEBI:35222	inhibitor	7	7	2113
100	VBS.inity	GO:0007340	acrosome_reaction	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	PRP.inhibitor‖of	GO:0030168	platelet_activation	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	VBS.reduc	CHEBI:27698	vanadium_atom	HYPONYM	CHEBI:35222	inhibitor	5	5	2113
100	PRP.inhibitor‖of	PR:000027718	RDR_large_chain*	HYPONYM	CHEBI:35222	inhibitor	6	6	2113
100	VBS.inhibit	CHEBI:26523	reactive_oxygen_species	VBS.have	PR:000009019	protein_IMPACT	5	5	361
100	HYPONYM	CHEBI:101278	diltiazem	HYPONYM	CHEBI:23888	drug	6	6	1222
100	VBS.decreas	GO:0007610	behavior	HYPONYM	CHEBI:23888	drug	5	5	1222
100	VBS.reduc	GO:0003824	catalytic_activity	HYPONYM	CHEBI:35222	inhibitor	9	9	2113

Table 25: Apriori results with confidence = 100%, excluding NP2s. The lack of diversity and the over-generality in the outer terms is evident
**ribonucleoside-diphosphate_reductase_large_chain* was abbreviated to fit in the available width

7.2.2 *Issues with Confidence as a metric*

It can be seen in Table 21, that the highest-confidence results have a high proportion of very general terms represented as their Outer property. While this is likely to be useful for some purposes, it has two main problems.

1. Rare inner terms and frequent outer terms are likely to yield a significant false positive rate. While we can try and deal with this crudely by specifying minimum support or minimum size for inner term, we would like something that takes both the inner and outer set sizes into account to improve our confidence in the results.

2. The rules that we are likely to find, are, at some level, *unsurprising*. If we are trying to produce a large list of assertions to inform, say, a machine-reasoning system that is starting from a position of no background knowledge, this may be useful. On the other hand, if we are trying to inform a human, these are not the highest-value terms to us.

With regard to these problems, we can take advantage of the additional rules in the Borgelt implementation of Apriori. Problem 1 is easiest to quantify, and the appropriate tool in this case is the *p*-value estimator*, which takes into account both the prevalence of the inner and outer terms as well as their overlap. Table 26 contains some of the lowest *p*-values, and it is obvious that the outer terms are much less general than in Table 24 and Table 25 (the rightmost column of the tables contain the number of chemicals described by the outer terms). Problem 2 we shall revisit later.

* Based on the χ^2 statistic; see `http://www.borgelt.net/doc/apriori/apriori.html`

log(*p*)	Inner Property			Outer Property			Together	Separately	
-999	HYPONYM	CHEBI:11814	3-hydroxy-3-methylglutaryl-CoA	HYPONYM	CHEBI:35664	HMG-CoA_reductase_inhibitor*	7	8	17
-999	HYPONYM	CHEBI:15756	palmitic_acid	HYPONYM	CHEBI:26607	saturated_fatty_acid	5	5	16
-999	HYPONYM	CHEBI:16814	dehydroepiandrosterone_sulfate	HYPONYM	CHEBI:17026	progesterone	4	5	13
-999	HYPONYM	CHEBI:16814	dehydroepiandrosterone_sulfate	HYPONYM	CHEBI:28689	dehydroepiandrosterone	4	5	11
-999	HYPONYM	CHEBI:16990	bilirubin	$\overbrace{\text{VBS.have}}$	NCBITaxon:10116	Rattus_norvegicus	4	5	16
-999	HYPONYM	CHEBI:16990	bilirubin	$\overbrace{\text{VBS.increas}}$	CHEBI:24433	group	4	5	13
-999	HYPONYM	CHEBI:17517	phosphatidylglycerol	$\overbrace{\text{VBS.compris}}$	CHEBI:18059	lipid	4	5	17
-999	HYPONYM	CHEBI:17517	phosphatidylglycerol	$\overbrace{\text{VBS.include}}$	CHEBI:18059	lipid	4	5	16
-999	HYPONYM	CHEBI:22693	barbiturates	HYPONYM	CHEBI:29745	barbiturate	4	5	15
-999	HYPONYM	CHEBI:23965	estradiol	HYPONYM	CHEBI:16469	17beta-estradiol	4	5	17
-999	HYPONYM	CHEBI:23965	estradiol	HYPONYM	CHEBI:17026	progesterone	4	5	13
-999	HYPONYM	CHEBI:26051	phosphoamino_acid	$\overbrace{\text{VBS.hydrolyz}}$	GO:0016791	phosphatase_activity	4	5	7
-999	HYPONYM	CHEBI:26271	proline	HYPONYM	CHEBI:28044	phenylalanine	4	5	11
-999	HYPONYM	CHEBI:26394	purine_nucleoside	HYPONYM	CHEBI:33838	nucleoside	11	13	48
-999	HYPONYM	CHEBI:26764	steroid_hormone	HYPONYM	CHEBI:24621	hormone	32	40	127
-999	HYPONYM	CHEBI:26831	N_N_-sulfonyldiurea	HYPONYM	CHEBI:35526	hypoglycemic_drug	7	8	31
-999	HYPONYM	CHEBI:27081	transition_element_atom	HYPONYM	CHEBI:33521	metal_atom	21	23	81
-999	HYPONYM	CHEBI:27314	water-soluble_vitamin	HYPONYM	CHEBI:33229	vitamin	11	12	56
-999	HYPONYM	CHEBI:28494	cardiolipin	HYPONYM	CHEBI:16038	phosphatidylethanolamine	4	5	9
-999	HYPONYM	CHEBI:28494	cardiolipin	$\overbrace{\text{VBS.compris}}$	CHEBI:18059	lipid	5	5	17
-999	HYPONYM	CHEBI:28494	cardiolipin	$\overbrace{\text{VBS.contain}}$	CHEBI:18059	lipid	5	5	21
-999	HYPONYM	CHEBI:28494	cardiolipin	$\overbrace{\text{VBS.include}}$	CHEBI:18059	lipid	5	5	16
-999	HYPONYM	CHEBI:28527	rutin	VBS.form	CHEBI:26519	radical	4	5	12
-999	HYPONYM	CHEBI:29995	aspartate_2-_	HYPONYM	CHEBI:27570	histidine	5	6	13
-999	HYPONYM	CHEBI:31941	oxaliplatin	PRP.iy_study‖of	CHEBI:24433	group	5	6	12
-999	HYPONYM	CHEBI:32460	cysteine_residue	HYPONYM	CHEBI:29917	thiol_group	4	5	10
-999	HYPONYM	CHEBI:33364	platinum	HYPONYM	CHEBI:33749	platinum_molecular_entity	5	6	10
-999	HYPONYM	CHEBI:33712	N-terminal_amino-acid_residue	HYPONYM	CHEBI:33708	amino-acid_residue	10	12	32
-999	HYPONYM	CHEBI:33712	N-terminal_amino-acid_residue	HYPONYM	CHEBI:33715	N-terminal-α-aar*	11	12	19
-999	HYPONYM	CHEBI:35338	amphetamines	HYPONYM	CHEBI:35337	central_nervous_system_stimulant	4	5	12

Table 26: Apriori results with low p values (a log(p) of -999 representing the minimum the software will represent), excluding NP2s. Compared to 25 the outer terms are much less general

**N-terminal_alpha-amino-acid_residue*, *hydroxymethylglutaryl-CoA_reductase_inhibitor*, were abbreviated to fit in the available width.

7.2.3 *Pre-processing the input*

We have the option of reducing the tautological relations discovered by excluding *all* NP2s — this has been done for several of the result sets above. It is extremely easy to implement but possibly more wasteful of data than we would like.

Given the information about associations between relations (Section 6.5), we could conflate those with a similarity score where we are reasonably sure that they are synonymous.

We can exclude properties where the relation doesn't appear to match the entity type (see Section 6.4).

We can also employ some of the heuristics for hypernyms described in chapter 5 — while some of these will not be applicable to other relation types, filtering by length is still likely to help in removing mis-resolved acronyms.

A further option is to exclude triples not mentioned in at least 2 (or n) papers, as discussed in 10 in Chapter 6.

7.2.4 *Post-processing the output*

Returning to the question touched on in Section 7.2.2 of making our rule-set more interesting to humans, *i.e.* containing fewer trivial relations, we have several possible options. Using a metric such as P-value will certainly help in this regard, since it is lower for pairs with outer terms that are too general — for these outer terms there is too high a probability than any given inner term would fall within them by chance.

Another technique would be to employ lexical matching to predict which rules a human reader might find trivial. One might start with simply looking for long substrings shared between the canonical forms of the inner and outer entities so that, for example, the rule (`is_a` water-soluble_vitamin ⇒ `is_a` vitamin) is detected as being obvious. For a more sophisticated approach, one might look to the type of lexical matching used in Bada and Hunter(2007)[13].

Relatedly, for rules where the inner and outer terms share the same relation, we can look at whether there is a taxonomic connection between the entities within our existing ontologies.

7.3 HUMAN ANNOTATION

It would be desirable to find a measure not only of whether one property is strictly entailed by another, but also when a property *typically* implies another. For instance [`is_a` nucleotide base analogue] will often imply [`is_a` mutagen], but there is no guarantee that this is the case (as compared to, for instance, [contains iron] implying [contains metal atom]).

A possibly useful property is *oddness*. Cruse(1986)[38]* says of oddness:

> One of the simplest and most basic semantic judgements one can make concerning an utterance in one's native language is whether it is to some degree odd or not.[38, p. 11]

> For the diagnosis of expected, possible, and unexpected traits, the *but*-test is extremely useful. This utilises the normality or abnormality of the form *P, but Q*. Consider the status of "can bark" as a trait of *dog*. First of all, "It's a dog" does not entail "It can bark" (since a dog may have a congenital malformation of the larynx, or some such); hence "can bark" is not a criterial trait. However, the following two sentences show it to be an expected trait:
>
> 1. It's a dog, but it can bark (odd)
> 2. It's a dog, but it can't bark (normal)†
>
> The sort of oddness exhibited by 1. may be termed **expressive paradox** since the expressive meaning carried by *but* is inappropriately deployed[38, p. 17]

We shall use this principle to identify pairs of properties where the converse of the implication is not *odd* in a but-test.

For instance *It's a β-lactam but it's not an antibiotic* is not odd, since an *expected* trait of a β-lactam molecule is that it is an antibiotic, even though of all the potential molecules that contain a beta-lactam ring a substantial proportion will be large enough that they are not liable to be taken up by bacteria. The basic ruleset used for annotation was defined by: minimum support 4, minimum confidence = 30%, maximum P-value = 50%. Certain steps were taken to exclude over-general and tautologous rules.

A rule was excluded if:

---

* Citing Bendix(1966)[14]

† These sentences are numbered 28 and 29 in Cruse; they are renumbered here for clarity.

- Either the head or the body of the rule had an entity corresponding to frequently mis-resolved terms such as PR:000009019 "Protein IMPACT", a frequent misresolution of "*impact*".
- Either the head or the body of the rule corresponded to certain overgeneral properties such as [`is_a` inhibitor].
- The canonical form of the entity in the head was found within the canonical form of the entity in the body, or vice versa, such as "*acid*" and "*amino acid*".
- The relation of the head — if the head relation was an NP2 — was found within the canonical form of the entity in the body, or vice versa, such as NP2.drug and "*anticancer drug*".

The first of these categories will tend to exclude false pairs; the last three will tend to exclude pairs that are true but are likely to be obvious or deducible by other means. Table 27 and Figure 29 show the effects of these exclusions on the number of pairs extracted at different thresholds for P and confidence. It is seen that thresholding by confidence produces sharp drops representing "popular" ratios such as ½ and ¾.

log(P)	Total	A	B	C
< -650	35235	34706	31467	26877
< -600	41458	40802	36909	31742
< -500	47482	46712	42103	36347
< -400	54954	54017	48553	42156
< -300	67656	66307	59280	51897
< -200	86061	83892	74526	65940
< -100	123496	117805	103952	93373
< -50	168497	155790	137122	124929
< -10	250828	209277	181941	168304
< 0	277697	213914	184523	170844
< 10	282521	213914	184523	170844
Confidence %	**Total**	**A**	**B**	**C**
100	6714	2820	2553	1596
> 90	7244	2930	2656	1648
> 80	14041	5592	4961	3522
> 70	30594	15044	13147	10828
> 60	49248	25965	22432	19095
> 50	136115	94657	81910	74808
> 40	206768	151669	130411	120073
> 30	282521	213914	184523	170844

Table 27: Numbers of inferred pairs of properties at various levels of confidence.
A: Excluding "over-general" properties such as [is_a molecule].
B: Excluding over-general properties and frequently mis-resolved entities such as *Protein IMPACT*.
C: Excluding obvious properties, frequently mis-resolved entities, and pairs where the canonical form of the outer term is found within the inner term, or vice versa.
For this analysis, restrictions other than that specified in the first column were not imposed.

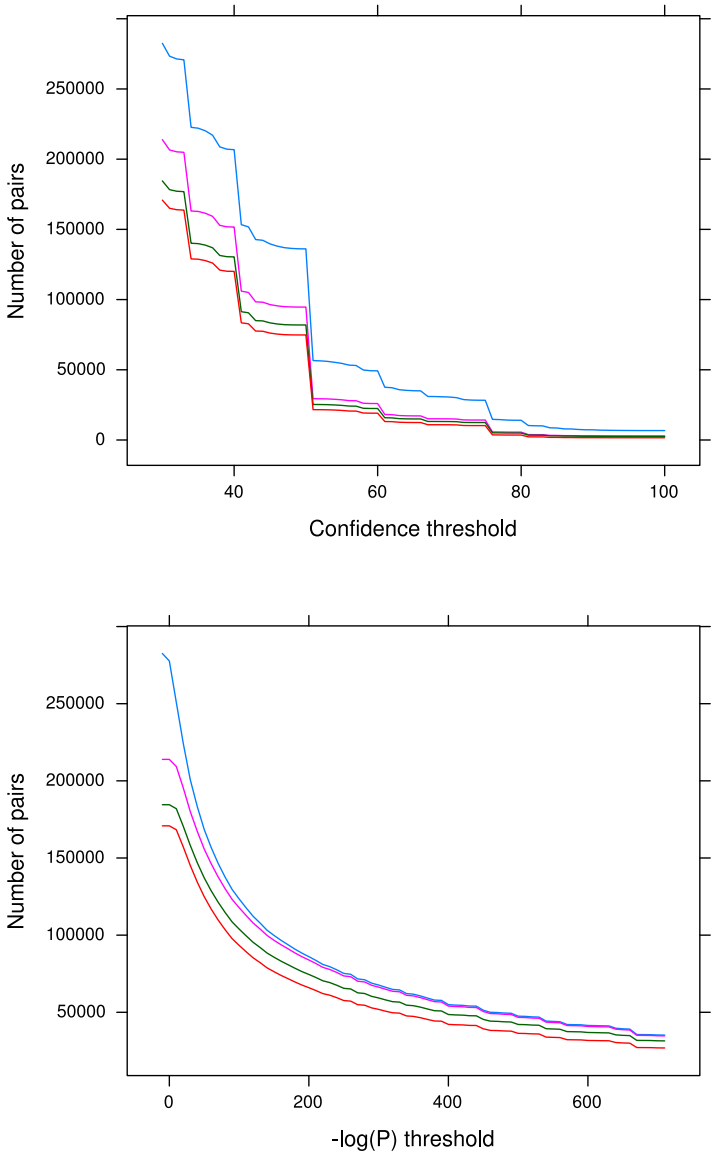


Figure 29: The effect of thresholds on number of pairs. This is a graphical representation of the data in Table 27. The series in both graphs, from top to bottom, correspond to: the total number of pairs; pairs excluding "over-general" properties; pairs excluding over-general properties and frequently mis-resolved entities; pairs excluding obvious properties, frequently mis-resolved entities, and pairs where the canonical form of the outer term is found within the inner term, or vice versa.

The rules were identified with Apriori, set with minimum rule support of 4, minimum confidence of 50%, and minimum P-value of 50%. A set of rules was sampled from across all P-values, shuffled, and supplied to the annotators. In addition to applying the oddness test to pairs of sentences, the annotators were asked to rate assertions directly from 1 to 5, according to the scale shown in Section 7.3.2 and with reference to the examples there. The annotators were both PhD-level chemists with a great deal of experience in annotation experiments.

7.3.1 *Annotation Guidelines*

Each assertion will be presented as a set of sentences of the form:

- This chemical Xs but it Ys
- This chemical Xs but it does not Y

You will be asked to rate whether the use of "*but*" in each sentence is "odd" in the way that the sentence "*This is a dog but it has four legs*" is odd. In this respect "*This is a dog but it has five legs*" is not odd (the dog is odd, the sentence is normal).

It is possible for both sentences to be odd:

- This is a car but it is blue
- This is a car but it is not blue

Assume that each sentence is not presented in any context beyond general biology.

The sentences are automatically generated. This means that some parts of the sentence may be incorrectly inflected; please ignore these mistakes when forming a judgement; a statement that *X haves an Y* should be interpreted as *X has a Y* and judged accordingly. In particular there will be statements of the form "*This chemical is (an) X but it is not Zed by (a) Y*" that are prone to this. The bracketed articles should be ignored (or not) as needed to make the sentence read smoothly.

The verbs have been stemmed, so rece → receive; modul → modulate; us → use etc.

- If a sentence's use of "*but*" is odd, mark it as Y

- If a sentence's use of "*but*" is not odd, mark it as N
- If a sentence is so difficult to understand that you are unable to form a judgement, mark it as X. This is for sentences that are somehow mangled, not for cases where it is hard to decide between Y and N.

7.3.2 *Direct assessment*

Please annotate the assertions of the form Property X → Property Y as below according to how the assertions are true.

1. Always or very often
2. Often enough to be relevant
3. Always or often, but only if construed in reverse direction
4. not often or never, but properties have plausible biological/chemical connection
5. No obvious connection or sentence not interpretable

Some examples:

1

`is_a` insecticide → `is_a` pesticide
`is_a` insecticide → kills insects

2

`is_a` organophosphate → `is_a` insecticide
`is_a` platinum compound → treats cancer

3

`is_a` insecticide → `is_a` organophosphate
treats cancer → `is_a` platinum compound

4

`is_a` antibiotic → `is_a` case of tuberculosis
`is_a` level of sex hormone → `is_a` estrogen receptor

5

`is_a` dopamine → is-had-by protein
`is_a` nobelium → `is_a` nitrogen oxide (even though you can see how the misresolution happens)

7.3.3 *Consultation*

The annotators should not make decisions on specific cases by comparing notes with each other or with the same third party. Consultation with others or use of reference resources is fine.

7.3.4 *Results*

The results, as confusion matrices, are as summarised in Table 28 and Table 30. Table 31 shows precisions and interannotator agreements for various criteria: the numeric scale annotations and the oddness test. We see that depending on which criterion was being applied, between 13% and 86% of the pairs were considered to be correct. 61% of pairs were judged to be correct by the standards of acheiving an *NY* on the Oddness test — that is, the first sentence of the pair was not considered odd, but the second sentence was, indicating that it is considered "normal" for a chemical which is described by the outer term to also be described by the inner term.

We find that the κ scores are rather lower than in the previous experiments, which hints towards a more subjective classification. The stricter annotation 1 has a somewhat better κ than that obtained if those propositions annotated as 2 are included.

	NN	NY	XX	YN	YY
NN	0	36	17	10	23
NY	1	111	6	13	14
XX	0	0	0	0	0
YN	0	4	0	2	2
YY	0	2	2	3	0

Table 28: Confusion matrix between annotators, using oddness measure. The first letter of each two-letter code is the judgement as to whether the first sentence "A chemical *Y*s but it does not *X*" is odd. The second letter is the judgement as to whether the second sentence "A chemical *Y*s but it *X*s" is odd. "X" indicates that the annotator was unable to form a judgement.

	BAD	GOOD
BAD	60	43
GOOD	34	111

Table 29: Confusion matrix between annotators. GOOD is NY; BAD is anything else

	1	2	3	4	5
1	14	3	1	3	0
2	15	23	9	18	0
3	0	4	3	1	1
4	15	29	13	48	5
5	1	4	1	16	21

Table 30: Confusion matrix between annotators, using numeric scale

Criterion	A1	A2	IAA	Precision	κ
1	21	45	210	0.133	0.349
1–2	86	108	164	0.391	0.295
1–3	95	135	162	0.464	0.321
1–4	205	221	220	0.859	0.538
NY	145	155	170	0.610	0.362

Table 31: Annotators' judgements of the precision of the extracted tuples according to different criteria. The first four rows summarise tuples given a numerical rating (on the scale 1 – 5) within the indicated range; the fifth row gives the equivalent numbers for tuples rated NY on the oddness measure. The columns show the number of tuples judged to be correct under each criterion by each annotator, the number of tuples for which the annotators agree, the precision of the tuples meeting the criterion (as judged by the mean of the two annotators' judgements), and Cohen's κ.

The "oddness" test yields a κ slightly higher than that obtained by just counting as correct propositions annotated as 1, but the two figures are close enough that they are best regarded as not significantly different, since even one more or fewer agreements between the annotators would reverse these figures. The best we can say from this is that the oddness test appears to be a potentially useful measure for identifying implications that are less limited than category 1 (*X always implies Y*) but less vague than category 4 (*X has something to do with Y*).

Something that should be examined is the use of "slim" ontologies to further reduce the sparsity of the dataset, by "rolling up" terms that only occur a small number of times to parent terms. This needs to be used with caution, though, as some assertions that are true with a child term will become false with a parent term. In particular, for hypernymic tuples, rolling up the hyponym is likely to yield a falsehood* (whereas rolling up the hypernym is unlikely to do so†, assuming the tuple was correct to begin with). The propagation methodology described in Bada and Hunter(2007)[13] would be very useful in this respect.

* *e.g.* [is_a cephalosporin ⇒ is_a antibiotic] does *not* imply [is_a heterocyclic compound ⇒ is_a antibiotic]

† *e.g.* [is_a cephalosporin ⇒ is_a antibiotic] implies [is_a cephalosporin ⇒ is_a antimicrobial]

Name	Examples
Relation type	HYPONYM; NP2 ; PRP
Relation	is_a ; VBS.inhibit ; PRP.modulator‖of
Entity	CHEBI:35338 ; GO:0042311 ; vasodilation
Property	is_a antibiotic ; VBS.inhibit amylase
Triple	[CHEBI:15365, VBS.reduce, GO:0042311]
Rule	[is_a CHEBI:35338 ⇒ VBS.reduce, GO:0042311]

Table 32: Guide to nomenclature conventions

7.4 SUMMARY

Combining entities and relations into properties allows us to find pairs of properties that are shared by more chemical entities than we would expect by chance. Depending on how we look for these pairs, we can identify potential synonyms, definitions, or implications. An annotation experiment showed that, depending on the criteria used, between 13% and 86% of the pairs discovered were potentially useful.

8

CONCLUSIONS

I have extracted a set of hypernyms of chemicals contained within the ChEBI ontology from parsed biomedical text after a named-entity recognition stage (the development of which is detailed in Chapter 4), using lexicosyntactic patterns. An annotation stage both allowed assessment of the results and provided training data based on which a classifier could be produced to further filter the results.

Further lexicosyntactic patterns were developed in Chapter 6 to capture relations between chemicals and other biological entities, without restricting their scope to a predetermined set of interaction types. I also investigated various means by which these relations can be characterized, including detecting possible synonyms on the basis of shared pairs of relations.

Combining entities and relations into properties allows us to find pairs of these properties that are shared by several chemical species, suggesting that there may be some link between them. Different criteria for extracting these pairs can identify potential synonyms, definitions, or implications. Annotation suggests that depending on the standards used, between 13% and 86% of these pairs are potentially useful.

8.1 CONTRIBUTIONS

8.1.1 *Theoretical contributions*

In Chapter 5 I have demonstrated a system to extract chemical hypernyms specific to the ChEBI ontology, making use of an NER stage and a parsing stage to allow accurate identification of complicated hypernyms that may appear in the text at a significant distance from the hypernyms they describe, and exploiting the phrase structure of the text to resolve hypernyms to ontology nodes as specifically as possible. I have also reviewed quantitatively and qualitatively a number of the factors that predispose toward errors, which may be informative for future development of such systems. Furthermore, I show how these factors can be integrated into a classification system that can make the precision of the output continuously variable, while still maintaining the link between each tuple and a sentence in the

biomedical literature. The analysis of errors within the case study in Appendix D, for example, is enabled by this link.

In Chapter 6 I extend the system to cover non-hypernymic relations between items not only from ChEBI but from other OBO ontologies. Rather than manually select relations corresponding to well-defined biological relations, the breadth of relations extracted suggests unsupervised ways of filtering and conflating relations based on the semantic types of their objects, and these approaches are explored. In a similar manner that noun phrases might provide suggestions to expand ontologies of entities, these patterns might provide an opportunity to expand such resources as the Relation Ontology [52].

In Chapter 7 a variety of techniques — Cosine similarity, PMI, and Apriori — are used to find properties either similar to or implied by other properties; this may help identify otherwise-undetected relations and synonyms, which could feed into the development or expansion of ontologies.

8.1.2 *Practical contributions*

This project has proved useful in helping the ChEBI curators make the project of assignments of edges in the ChEBI ontology both faster and more comprehensive. The data generated in Chapter 5 are being used to assist the ChEBI team in their curation activities. The data have been integrated into a curation tool (screenshots of which are in Appendix E) whereby, by searching for a particular ChEBI term, curators can easily view candidate hypernyms, alongside the abstracts that provide the evidence for the hypernymy. Because the hypernyms are already resolved to ChEBI terms, and because the evidence is attached, the edge between nodes in the ontology can be added with a single click, rather than necessarily requiring manual searching of the literature. ChEBI intends that the non-hypernymic triples from Chapter 6 be used in a similar way in the near future.

8.1.2.1 *ChEBI curatorial assessment*

The ChEBI curatorial team provided the following description and qualitative assessment of their use of the system.*

* It should be noted that the user interface features mentioned were implemented by Adriano Dekker and other EBI developers, and not by the author.

Description

The Data Extraction (DE) tool is now integrated into the curation tool used by all of the ChEBI curators to assist in the classification of entities both structurally (is_a triterpenoid, is_a carboxylic acid, etc.) and biologically (has_role anti-inflammatory drug; has_role cyclooxygenase inhibitor, etc.) within the ChEBI Ontology.

An essential step of the curation process for a particular entity is to search PubMed for citations that are relevant to that entity (e.g. review articles, first and improved syntheses, biological properties, etc.). The curation tool includes a 'citation finder' tool that compiles the names/synonyms stored in ChEBI for a particular entity into a suitable query for CiteXplore/PubMed, displaying the results in a tabular form that enables the curator to easily find desired citations and add them to the ChEBI database by a point-and-click process. The DE tool has been incorporated as a part of this curation page, so that it is automatically accessed whenever a curator click on the 'citation finder' tab.

Limitations

A feature of the DE tool is that it can only search for data about entities which were present in ChEBI when the DE indexes were last updated. This is a serious limitation, as it means that it cannot be used for to assist in the curation of entities which are new to ChEBI (ca. 95% of current curation). If we knew that (for example) indexing would be run overnight every night, then we could adjust the curation process to take this into account, creating a new entity one day and completing the entry the following day. Unfortunately, this is not currently the case, with the results that curators only curate on average perhaps one entry a week where the DE tool can be used.

When the DE tool can be used, some minor limitations are apparent — for example, it often suggests a more general term (has_role drug) when a more specific term (has_role non-steroidal anti-inflammatory drug) is already present, while its suggestions for structural classification are frequently absurd (e.g. is_a magnesium atom). These are minor limitations, however — adding a role which doesn't need to be added is unnecessary but not wrong and takes very little time, while silly structural classifications are usually very obvious to the curators, who are all chemists.

Despite the limitations listed, when it is used the DE tool is often very useful, finding (particularly) biological roles that would take curators ages to find (if they ever did) if they had to rely on manual searching of PubMed.

8.1.2.2 *Comments*

It is clear from the curators' comments that (subjectively at least) the software is performing usefully so far in the situations where it is applicable. Furthermore, most of the negative issues discussed should be reasonably straightforward to improve (once the parsed corpus is regenerated).

The application of the filtering techniques described in Section 5.5 but not yet integrated into the curatorial tools should help reduce some of the "absurd" cases (the 'magnesium' example is one of those discussed in Section 5.4.3.1. Regenerating the results with new additions to ChEBI should also not be problematic; since the NER stage was based on a machine-learning approach rather than dictionary-matching the existing ChEBI contents, it is expected that most new additions to ChEBI will already have been identified as being chemical species, thus repeating the NER (and subsequent parsing, since NER is a prerequisite for parsing as discussed in Section 4.4) is not required.

8.2 SUGGESTED FUTURE WORK

There are perhaps three main avenues down which the work could extend - improving recall and extending scope; improving precision, perhaps to fit the data set for other uses such as generating ontology structures without a curator-in-the-loop ; and analysing the datasets to extract new kinds of relations from them.

8.2.1 *Broader!*

- Broaden the scope to include triples not including chemicals — if we are identifying relations between ChEBI and other ontologies, it may well be desirable to identify relations between and within these ontologies.
- Additional ontologies — the architecture of the software makes it easy to add ontologies providing they have definitions available in OBO format. However, more entities will inevitably mean more "collisions" where entities in different ontologies share the same synonyms.
- Additional LSPs. There are specific patterns such as "X [verb] [preposition] Y" that might be included; more generally it may be possible to define looser patterns that match more widely, and investigate the ex-

tent to which these can later be filtered and conflated. Some patterns that involve more than one entity (in addition to the chemical entity) may be appropriate.

- Anaphor & reference resolution — this is a well-characterised problem that leads to relations being missed, or over-general entities identified when a phrase such as "the protein" is used in place of a more specific identifier.
- Inclusion of relations involving strings not resolvable to ontological entities — perhaps those matching a particular pattern equivalent to an NER step? Good way of finding new candidates for nodes — or new synonyms for existing nodes.

8.2.2 *Sharper!*

- Word-sense disambiguation beyond abbreviation detection. At a simple level, we might imagine identifying abbreviations that do not have an explicit expansion in the document, making a list of the candidates based on explicit expansions across the corpus, and using the expansion mentioned most frequently within the document.
- NER steps per semantic type - so, for instance, a protein requires being matched by a protein named entity recognizer. Would push down yield but might increase accuracy, especially in terms of reducing cross-domain misresolutions and (assuming the NER stage considered frequencies in general text as discussed in Chapter 4) resolutions such as identifying "Impact" and "Mice" as proteins.
- Identify redundant patterns — being able to better merge variants, whether spelling variations (*catalyse/catalyze*), synonyms (*increase/raise*), or syntactic variations (*inhibits X/inhibitor of X/X inhibitor*) should both yield a cleaner, more human-readable data set, and less sparsity in inference.
- As we can see in Appendix D, mis-parsing of lists as appositions is a major source of error. While improving this directly would probably require working with the developers of the syntactic parser, it is possible that looking at features such as length of list, apparent semantic types of appositions, and making sure that phrases taken to be appositions agree in terms of number, might give some basis for filtering out at least some of the false positives.

8.2.3 *Deeper!*

- Associate at a step remove; not just with such properties as [`inhibits` *Protein P*] but [`inhibits` *a protein annotated in GOA with GO term T*] - so we might hope that while [`is_a` Neurotransmitter] might correlate with [VBS.inhibits(protein with GO Cellular Component annotation *synapse*)]
- Develop patterns that can be mapped to Entity–Quality descriptions[74] such as *Elevated* Na^+.
- Integrate existing ontological data into the inference step, in addition to that directly extracted from the text.
- It would be useful to add more context-discrimination to assertions about behaviours of molecules. Some pairs of examples that it may be possible and desirable to distinguish between: whether an effect is observed *in vivo* or *in vitro*; in "normal" state or in a perturbed state (disease state or experimental protocol); in prokaryotes or in eukaryotes.

FINIS

A

FREQUENCIES OF SEMANTIC TYPES

Table 33 shows the semantic profiles of various relation types. Each row is a relation type. Each column shows the frequency of different types of arguments.

For instance, for the relationship `VBS.modify` and the column `PROT`, the cell answers the question: *For relationships of the form `[Chemical X] modifies [Entity Y]`, how many **distinct** `Y`s were proteins, and what proportion does that represent of all the distinct `Y` (of all semantic types)?* The relations are grouped by those having a preference for each semantic type (in the same order as the columns) - so for example those where `Y` is predominantly a ChEBI entity come first, followed by Disease Ontology, GO Biological Process, GO Cellular Component, GO Molecular Function, Species ("Taxon"), Protein, and any others.

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
Relations that have a preference for arguments of type: CHEBI									
NP2.deriv	342 (**97%**)	0	0	2 (0%)	0	0	4 (1%)	2 (0%)	350
PRP.analog‖of	194 (**96%**)	0	0	0	0	0	8 (3%)	0	202
PRP.class‖of	85 (**95%**)	0	0	1 (1%)	1 (1%)	0	1 (1%)	1 (1%)	89
VBS.substitut	78 (**95%**)	0	0	0	0	0	4 (4%)	0	82
NP2.donor	50 (**94%**)	0	0	0	1 (1%)	0	2 (3%)	0	53
NP2.analog	364 (**93%**)	0	3 (0%)	2 (0%)	0	1 (0%)	16 (4%)	2 (0%)	388

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
PRP.deriv‖of	117 (93%)	0	1 (0%)	1 (0%)	1 (0%)	0	5 (4%)	0	125
PRP.group‖of	67 (93%)	0	0	0	0	1 (1%)	4 (5%)	0	72
VBS.replac	250 (92%)	0	0	0	1 (0%)	0	16 (5%)	3 (1%)	270
PRP.amount‖of	49 (92%)	0	0	0	0	0	3 (5%)	1 (1%)	53
PRP.effect‖of	220 (92%)	3 (1%)	3 (1%)	0	0	0	11 (4%)	1 (0%)	238
PRP.concentr‖of	81 (92%)	0	0	0	0	0	6 (6%)	1 (1%)	88
VBS.remain	61 (91%)	1 (1%)	0	1 (1%)	0	0	4 (5%)	0	67
VBS.displac	68 (90%)	0	0	1 (1%)	1 (1%)	0	5 (6%)	0	75
VBS.bear	65 (90%)	0	0	3 (4%)	0	0	3 (4%)	1 (1%)	72
$\overline{\text{VBS.replac}}$	99 (89%)	0	1 (0%)	0	0	0	11 (9%)	0	111
NP2.antibiot	49 (89%)	0	0	0	3 (5%)	0	2 (3%)	1 (1%)	55
NP2.precursor	145 (88%)	0	6 (3%)	4 (2%)	0	0	6 (3%)	2 (1%)	163
PRP.type‖of	63 (88%)	1 (1%)	1 (1%)	2 (2%)	1 (1%)	0	3 (4%)	0	71
PRP.metabolit‖of	222 (88%)	2 (0%)	14 (5%)	0	0	0	13 (5%)	0	251
VBS.carry	61 (88%)	0	1 (1%)	0	0	0	4 (5%)	3 (4%)	69
PRP.form‖of	189 (87%)	5 (2%)	3 (1%)	0	2 (0%)	0	15 (6%)	1 (0%)	215
PRP.sourc‖of	142 (87%)	3 (1%)	5 (3%)	0	1 (0%)	0	10 (6%)	1 (0%)	162
VBS.becom	85 (87%)	1 (1%)	0	2 (2%)	0	0	7 (7%)	2 (2%)	97
NP2.compound	233 (87%)	0	7 (2%)	6 (2%)	4 (1%)	0	15 (5%)	1 (0%)	266
NP2.concentr	67 (85%)	0	0	3 (3%)	0	0	8 (10%)	0	78
PRP.precursor‖of	166 (85%)	0	21 (10%)	1 (0%)	1 (0%)	0	5 (2%)	0	194

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
NP2.metabolit	182 (**85%**)	0	4 (1%)	7 (3%)	1 (0%)	1 (0%)	16 (7%)	2 (0%)	213
NP2.specy	54 (**84%**)	1 (1%)	0	4 (6%)	1 (1%)	0	3 (4%)	1 (1%)	64
VBS.remov	46 (**83%**)	0	0	3 (5%)	0	0	6 (10%)	0	55
NP2.drug	172 (**83%**)	11 (5%)	2 (0%)	2 (0%)	9 (4%)	0	8 (3%)	2 (0%)	206
NP2.effect	55 (**83%**)	2 (3%)	1 (1%)	2 (3%)	0	0	6 (9%)	0	66
VBS.yield	54 (**83%**)	0	0	4 (6%)	1 (1%)	0	3 (4%)	3 (4%)	65
VBS.contain	423 (**82%**)	1 (0%)	1 (0%)	11 (2%)	7 (1%)	1 (0%)	58 (11%)	8 (1%)	510
NP2.agent	173 (**82%**)	4 (1%)	12 (5%)	4 (1%)	10 (4%)	0	4 (1%)	2 (0%)	209
$\overbrace{\text{VBS.include}}$	159 (**81%**)	5 (2%)	5 (2%)	4 (2%)	1 (0%)	0	20 (10%)	2 (1%)	196
VBS.oxidiz	51 (**80%**)	0	0	2 (3%)	1 (1%)	0	6 (9%)	3 (4%)	63
$\overbrace{\text{VBS.inhibit}}$	80 (**80%**)	0	3 (3%)	2 (2%)	0	0	14 (14%)	0	99
NP2.level	44 (**80%**)	0	2 (3%)	2 (3%)	1 (1%)	0	6 (10%)	0	55
PRP.%‖of	50 (**79%**)	1 (1%)	2 (3%)	3 (4%)	1 (1%)	1 (1%)	4 (6%)	1 (1%)	63
NP2.group	68 (**79%**)	2 (2%)	1 (1%)	1 (1%)	2 (2%)	0	12 (13%)	0	86
VBS.be	1222 (**78%**)	20 (1%)	42 (2%)	23 (1%)	26 (1%)	2 (0%)	183 (11%)	34 (2%)	1552
VBS.include	59 (**78%**)	2 (2%)	1 (1%)	0	1 (1%)	1 (1%)	9 (12%)	2 (2%)	75
NP2.form	51 (**77%**)	0	4 (6%)	3 (4%)	0	0	7 (10%)	1 (1%)	66
NP2.substanc	44 (**77%**)	0	5 (8%)	2 (3%)	1 (1%)	0	4 (7%)	1 (1%)	57
$\overbrace{\text{VBS.increas}}$	123 (**75%**)	3 (1%)	8 (4%)	4 (2%)	3 (1%)	3 (1%)	19 (11%)	0	163
NP2.treatment	46 (**75%**)	10 (**16%**)	0	0	0	0	4 (6%)	1 (1%)	61
PRP.level‖of	42 (**73%**)	0	2 (3%)	0	1 (1%)	0	11 (**19%**)	1 (1%)	57

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
$\overline{\text{VBS.be}}$	696 (**73%**)	15 (1%)	36 (3%)	20 (2%)	19 (2%)	3 (0%)	132 (13%)	27 (2%)	948
VBS.repres	105 (**72%**)	1 (0%)	3 (2%)	5 (3%)	5 (3%)	0	23 (**15%**)	2 (1%)	144
$\overline{\text{VBS.form}}$	64 (**72%**)	0	3 (3%)	4 (4%)	1 (1%)	0	14 (**15%**)	2 (2%)	88
VBS.provid	71 (**72%**)	2 (2%)	6 (6%)	4 (4%)	0	0	14 (14%)	1 (1%)	98
HYPONYM	3247 (**72%**)	110 (2%)	176 (3%)	84 (1%)	107 (2%)	9 (0%)	678 (**15%**)	73 (1%)	4484
NP2.molecul	94 (**72%**)	0	9 (6%)	7 (5%)	4 (3%)	0	14 (10%)	2 (1%)	130
VBS.releas	54 (**72%**)	0	1 (1%)	3 (4%)	1 (1%)	0	14 (**18%**)	2 (2%)	75
VBS.give	46 (**71%**)	0	4 (6%)	1 (1%)	0	0	12 (**18%**)	1 (1%)	64
VBS.form	115 (**71%**)	0	1 (0%)	18 (11%)	1 (0%)	0	21 (13%)	4 (2%)	160
VBS.possess	37 (**71%**)	0	3 (5%)	3 (5%)	2 (3%)	0	6 (11%)	1 (1%)	52
$\overline{\text{VBS.decreas}}$	71 (**71%**)	1 (1%)	6 (6%)	5 (5%)	3 (3%)	1 (1%)	12 (12%)	1 (1%)	100
NP2.constitu	41 (**70%**)	0	0	11 (**18%**)	0	0	4 (6%)	2 (3%)	58
NP2.residu	43 (**70%**)	0	0	2 (3%)	4 (6%)	0	12 (**19%**)	0	61
VBS.gener	46 (**69%**)	1 (1%)	7 (10%)	3 (4%)	2 (3%)	0	6 (9%)	1 (1%)	66
$\overline{\text{VBS.reduc}}$	111 (**68%**)	1 (0%)	3 (1%)	7 (4%)	6 (3%)	2 (1%)	28 (**17%**)	4 (2%)	162
VBS.constitut	46 (**67%**)	0	1 (1%)	6 (8%)	0	0	11 (**16%**)	4 (5%)	68
$\overline{\text{VBS.affect}}$	39 (**67%**)	0	5 (8%)	1 (1%)	1 (1%)	1 (1%)	10 (**17%**)	1 (1%)	58
NP2.compon	84 (**66%**)	0	3 (2%)	23 (**18%**)	4 (3%)	0	10 (7%)	3 (2%)	127
NP2.receptor	48 (**64%**)	0	0	5 (6%)	3 (4%)	0	18 (**24%**)	1 (1%)	75
VBS.us	143 (**60%**)	2 (0%)	11 (4%)	8 (3%)	6 (2%)	3 (1%)	59 (**24%**)	6 (2%)	238
VBS.compris	30 (**58%**)	0	1 (1%)	7 (13%)	0	0	8 (**15%**)	5 (9%)	51

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.mimic	48 (**57%**)	4 (4%)	16 (**19%**)	4 (4%)	2 (2%)	0	7 (8%)	2 (2%)	83
NP2.product	80 (**55%**)	0	31 (**21%**)	4 (2%)	3 (2%)	0	24 (**16%**)	2 (1%)	144
VBS.have	148 (**54%**)	6 (2%)	21 (7%)	11 (4%)	11 (4%)	0	71 (**26%**)	4 (1%)	272
PRP.product‖of	170 (**53%**)	0	98 (**30%**)	2 (0%)	11 (3%)	0	35 (11%)	1 (0%)	317
PRP.constitu‖of	54 (**53%**)	0	5 (4%)	20 (**19%**)	1 (0%)	0	14 (13%)	7 (6%)	101
$\overbrace{\text{VBS.releas}}$	61 (**52%**)	1 (0%)	12 (10%)	11 (9%)	5 (4%)	1 (0%)	20 (**17%**)	5 (4%)	116
$\overbrace{\text{VBS.contain}}$	234 (**51%**)	9 (1%)	8 (1%)	55 (12%)	10 (2%)	3 (0%)	121 (**26%**)	18 (3%)	458
PRP.compon‖of	131 (**50%**)	3 (1%)	39 (**15%**)	51 (**19%**)	1 (0%)	1 (0%)	23 (8%)	10 (3%)	259
VBS.bind	58 (**49%**)	0	3 (2%)	7 (5%)	3 (2%)	0	38 (**32%**)	8 (6%)	117
NP2.substrat	71 (**49%**)	0	3 (2%)	3 (2%)	6 (4%)	0	59 (**40%**)	2 (1%)	144
NP2.activity	25 (**48%**)	0	5 (9%)	2 (3%)	6 (11%)	0	12 (**23%**)	2 (3%)	52
VBS.lower	31 (**46%**)	1 (1%)	11 (**16%**)	1 (1%)	7 (10%)	0	12 (**18%**)	3 (4%)	66
NP2.antagonist	110 (**46%**)	0	5 (2%)	2 (0%)	19 (8%)	0	100 (**42%**)	1 (0%)	237
PRP.precursor‖for	27 (**45%**)	0	26 (**44%**)	0	0	0	5 (8%)	1 (1%)	59
$\overbrace{\text{VBS.gener}}$	37 (**44%**)	1 (1%)	12 (14%)	5 (6%)	3 (3%)	1 (1%)	20 (**24%**)	4 (4%)	83
VBS.stabil	42 (**43%**)	3 (3%)	5 (5%)	23 (**23%**)	7 (7%)	0	15 (**15%**)	1 (1%)	96
NP2.ligand	52 (**43%**)	0	0	8 (6%)	9 (7%)	0	45 (**37%**)	6 (5%)	120
$\overbrace{\text{VBS.remov}}$	31 (**41%**)	1 (1%)	7 (9%)	3 (4%)	6 (8%)	0	22 (**29%**)	4 (5%)	74
$\overbrace{\text{VBS.have}}$	71 (**41%**)	4 (2%)	3 (1%)	15 (8%)	2 (1%)	7 (4%)	62 (**35%**)	9 (5%)	173
$\overbrace{\text{VBS.us}}$	68 (**40%**)	7 (4%)	19 (11%)	9 (5%)	5 (2%)	4 (2%)	47 (**27%**)	10 (5%)	169
VBS.produc	80 (**40%**)	30 (**15%**)	48 (**24%**)	4 (2%)	3 (1%)	1 (0%)	28 (14%)	5 (2%)	199

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
$\overleftarrow{\text{VBS.provid}}$	22 (**40%**)	0	8 (14%)	8 (14%)	1 (1%)	0	16 (**29%**)	0	55
$\overleftarrow{\text{VBS.lack}}$	22 (**39%**)	2 (3%)	2 (3%)	8 (14%)	0	2 (3%)	15 (**26%**)	5 (8%)	56
VBS.protect	43 (**39%**)	4 (3%)	10 (9%)	17 (**15%**)	3 (2%)	4 (3%)	23 (**20%**)	6 (5%)	110
NP2.agonist	68 (**38%**)	0	1 (0%)	6 (3%)	19 (10%)	0	82 (**46%**)	0	176
$\overleftarrow{\text{VBS.bind}}$	79 (**38%**)	0	2 (0%)	14 (6%)	11 (5%)	2 (0%)	90 (**43%**)	9 (4%)	207
$\overleftarrow{\text{VBS.produc}}$	60 (**37%**)	6 (3%)	20 (12%)	10 (6%)	7 (4%)	4 (2%)	48 (**30%**)	4 (2%)	159
VBS.modify	57 (**37%**)	5 (3%)	54 (**35%**)	6 (3%)	7 (4%)	0	21 (13%)	4 (2%)	154
PRP.determin‖of	27 (**36%**)	7 (9%)	26 (**35%**)	2 (2%)	4 (5%)	0	7 (9%)	1 (1%)	74
VBS.show	35 (**34%**)	5 (4%)	28 (**27%**)	3 (2%)	6 (5%)	0	20 (**19%**)	4 (3%)	101
$\overleftarrow{\text{VBS.show}}$	18 (**33%**)	0	3 (5%)	5 (9%)	1 (1%)	3 (5%)	21 (**39%**)	2 (3%)	53
PRP.predictor‖of	18 (**33%**)	22 (**41%**)	7 (13%)	1 (1%)	0	0	5 (9%)	0	53
$\overleftarrow{\text{VBS.rece}}$	27 (**33%**)	7 (8%)	3 (3%)	5 (6%)	0	7 (8%)	31 (**38%**)	0	80
PRP.antagonist‖of	27 (**33%**)	1 (1%)	5 (6%)	2 (2%)	11 (13%)	0	34 (**41%**)	1 (1%)	81
VBS.reduc	156 (**32%**)	33 (6%)	144 (**30%**)	10 (2%)	42 (8%)	0	72 (**15%**)	18 (3%)	475
VBS.reach	19 (**32%**)	3 (5%)	1 (1%)	20 (**34%**)	0	0	14 (**24%**)	1 (1%)	58
VBS.antagon	25 (**32%**)	5 (6%)	24 (**31%**)	2 (2%)	6 (7%)	0	13 (**16%**)	2 (2%)	77
VBS.maintain	17 (**32%**)	1 (1%)	19 (**35%**)	8 (**15%**)	4 (7%)	0	4 (7%)	0	53
VBS.exhibit	22 (**29%**)	1 (1%)	21 (**28%**)	2 (2%)	7 (9%)	0	19 (**25%**)	2 (2%)	74
VBS.decreas	96 (**27%**)	18 (5%)	102 (**29%**)	7 (2%)	42 (12%)	0	72 (**20%**)	13 (3%)	350
$\overleftarrow{\text{VBS.convert}}$	23 (**26%**)	0	6 (6%)	4 (4%)	7 (8%)	1 (1%)	45 (**51%**)	1 (1%)	87
VBS.chang	20 (**25%**)	2 (2%)	29 (**37%**)	4 (5%)	8 (10%)	0	13 (**16%**)	1 (1%)	77

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.inactiv	17 (**25%**)	0	7 (10%)	3 (4%)	11 (**16%**)	0	27 (**39%**)	3 (4%)	68
VBS.target	19 (**24%**)	1 (1%)	9 (11%)	15 (**19%**)	7 (9%)	0	21 (**27%**)	5 (6%)	77
VBS.increas	117 (**24%**)	7 (1%)	148 (**30%**)	17 (3%)	56 (11%)	1 (0%)	118 (**24%**)	17 (3%)	481
VBS.block	75 (**23%**)	4 (1%)	129 (**39%**)	14 (4%)	28 (8%)	0	62 (**19%**)	13 (4%)	325
PRP.ligand‖for	15 (**22%**)	0	2 (2%)	2 (2%)	5 (7%)	1 (1%)	41 (**61%**)	1 (1%)	67
VBS.abolish	23 (**22%**)	5 (4%)	48 (**46%**)	3 (2%)	8 (7%)	0	12 (11%)	5 (4%)	104
VBS.alter	43 (**21%**)	4 (2%)	90 (**45%**)	12 (6%)	20 (10%)	0	24 (12%)	3 (1%)	196
NP2.blocker	30 (**21%**)	0	29 (**20%**)	4 (2%)	24 (**17%**)	0	50 (**35%**)	4 (2%)	141
VBS.restor	16 (**20%**)	0	36 (**46%**)	0	15 (**19%**)	0	7 (9%)	3 (3%)	77
NP2.protein	20 (**20%**)	1 (1%)	7 (7%)	10 (10%)	20 (**20%**)	2 (2%)	36 (**37%**)	1 (1%)	97
$\widehat{\text{VBS.hydrolyz}}$	13 (**19%**)	0	0	6 (9%)	8 (12%)	0	34 (**51%**)	5 (7%)	66
$\widehat{\text{VBS.requir}}$	27 (**19%**)	5 (3%)	38 (**27%**)	12 (8%)	19 (13%)	3 (2%)	27 (**19%**)	8 (5%)	139
PRP.agent‖for	11 (**18%**)	34 (**58%**)	9 (**15%**)	2 (3%)	0	0	1 (1%)	1 (1%)	58
VBS.affect	87 (**18%**)	9 (1%)	199 (**43%**)	29 (6%)	47 (10%)	1 (0%)	73 (**15%**)	15 (3%)	460
VBS.revers	22 (**18%**)	18 (**15%**)	49 (**41%**)	2 (1%)	6 (5%)	1 (0%)	17 (14%)	3 (2%)	118
$\widehat{\text{VBS.transport}}$	10 (**18%**)	0	2 (3%)	4 (7%)	5 (9%)	0	30 (**55%**)	3 (5%)	54
NP2.activat	14 (**18%**)	0	5 (6%)	7 (9%)	11 (14%)	0	39 (**51%**)	0	76
VBS.facilit	12 (**18%**)	1 (1%)	42 (**63%**)	1 (1%)	2 (3%)	0	5 (7%)	3 (4%)	66
VBS.prevent	39 (**17%**)	52 (**23%**)	92 (**42%**)	7 (3%)	6 (2%)	1 (0%)	14 (6%)	6 (2%)	217
PRP.agent‖in	10 (**17%**)	38 (**67%**)	1 (1%)	1 (1%)	0	2 (3%)	4 (7%)	0	56
VBS.attenu	25 (**17%**)	21 (14%)	70 (**48%**)	4 (2%)	6 (4%)	0	12 (8%)	6 (4%)	144

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.inity	9 (**17%**)	4 (7%)	31 (**59%**)	1 (1%)	2 (3%)	0	3 (5%)	2 (3%)	52
VBS.influenc	32 (**17%**)	3 (1%)	95 (**51%**)	7 (3%)	20 (10%)	0	24 (12%)	5 (2%)	186
VBS.suppress	44 (**15%**)	20 (7%)	139 (**50%**)	2 (0%)	23 (8%)	0	43 (**15%**)	5 (1%)	276
VBS.enter	11 (**15%**)	2 (2%)	10 (14%)	31 (**44%**)	0	0	12 (**17%**)	3 (4%)	69
VBS.augment	11 (**15%**)	2 (2%)	35 (**50%**)	2 (2%)	8 (11%)	0	9 (13%)	2 (2%)	69
VBS.diminish	9 (**15%**)	1 (1%)	31 (**54%**)	1 (1%)	8 (14%)	0	6 (10%)	1 (1%)	57
PRP.drug‖in	9 (**15%**)	40 (**70%**)	2 (3%)	0	0	1 (1%)	5 (8%)	0	57
VBS.enhanc	47 (**15%**)	8 (2%)	151 (**50%**)	5 (1%)	52 (**17%**)	0	25 (8%)	11 (3%)	299
NP2.inhibitor	86 (**15%**)	0	131 (**23%**)	27 (4%)	48 (8%)	0	254 (**44%**)	23 (4%)	569
Relations that have a preference for arguments of type: DISEASE									
PRP.treatment‖for	8 (3%)	238 (**89%**)	4 (1%)	0	0	0	13 (4%)	2 (0%)	265
PRP.treatment‖of	4 (5%)	69 (**89%**)	2 (2%)	0	0	0	1 (1%)	1 (1%)	77
PRP.therapy‖for	4 (3%)	87 (**86%**)	1 (0%)	1 (0%)	0	0	7 (6%)	1 (0%)	101
PRP.caus‖of	3 (3%)	73 (**82%**)	11 (12%)	0	0	0	1 (1%)	1 (1%)	89
PRP.risk␣factor‖for	3 (4%)	49 (**80%**)	5 (8%)	0	0	0	3 (4%)	1 (1%)	61
PRP.drug‖for	4 (5%)	63 (**79%**)	3 (3%)	2 (2%)	0	0	6 (7%)	1 (1%)	79
VBS.treat	7 (11%)	44 (**72%**)	1 (1%)	1 (1%)	1 (1%)	2 (3%)	4 (6%)	1 (1%)	61
PRP.drug‖in	9 (**15%**)	40 (**70%**)	2 (3%)	0	0	1 (1%)	5 (8%)	0	57
PRP.agent‖in	10 (**17%**)	38 (**67%**)	1 (1%)	1 (1%)	0	2 (3%)	4 (7%)	0	56
PRP.agent‖for	11 (**18%**)	34 (**58%**)	9 (**15%**)	2 (3%)	0	0	1 (1%)	1 (1%)	58
PRP.factor‖for	4 (3%)	56 (**55%**)	30 (**29%**)	1 (0%)	2 (1%)	1 (0%)	4 (3%)	3 (2%)	101

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
PRP.predictor‖of	18 (33%)	22 (41%)	7 (13%)	1 (1%)	0	0	5 (9%)	0	53
PRP.factor‖in	3 (4%)	26 (36%)	38 (52%)	1 (1%)	2 (2%)	0	2 (2%)	0	72
VBS.caus	35 (14%)	81 (33%)	79 (32%)	10 (4%)	3 (1%)	0	28 (11%)	5 (2%)	241
PRP.marker‖of	8 (14%)	18 (31%)	22 (38%)	3 (5%)	1 (1%)	0	4 (7%)	1 (1%)	57
VBS.improve	16 (12%)	33 (25%)	57 (43%)	2 (1%)	8 (6%)	0	13 (9%)	3 (2%)	132
VBS.prevent	39 (17%)	52 (23%)	92 (42%)	7 (3%)	6 (2%)	1 (0%)	14 (6%)	6 (2%)	217
VBS.control	4 (3%)	23 (22%)	52 (50%)	2 (1%)	5 (4%)	1 (0%)	12 (11%)	4 (3%)	103
VBS.induce	63 (10%)	118 (19%)	211 (34%)	16 (2%)	38 (6%)	2 (0%)	141 (22%)	26 (4%)	615
NP2.treatment	46 (75%)	10 (16%)	0	0	0	0	4 (6%)	1 (1%)	61
VBS.revers	22 (18%)	18 (15%)	49 (41%)	2 (1%)	6 (5%)	1 (0%)	17 (14%)	3 (2%)	118
VBS.produc	80 (40%)	30 (15%)	48 (24%)	4 (2%)	3 (1%)	1 (0%)	28 (14%)	5 (2%)	199
Relations that have a preference for arguments of type: GO_BIO_PROCESS									
VBS.promot	4 (2%)	5 (3%)	106 (74%)	2 (1%)	8 (5%)	1 (0%)	10 (7%)	6 (4%)	142
VBS.impair	5 (6%)	3 (4%)	55 (73%)	3 (4%)	4 (5%)	0	4 (5%)	1 (1%)	75
PRP.regul‖of	12 (6%)	1 (0%)	132 (66%)	5 (2%)	11 (5%)	0	35 (17%)	3 (1%)	199
VBS.trigger	7 (8%)	9 (10%)	55 (64%)	2 (2%)	1 (1%)	0	5 (5%)	6 (7%)	85
PRP.medy‖of	5 (5%)	14 (14%)	61 (64%)	1 (1%)	0	0	7 (7%)	7 (7%)	95
VBS.facilit	12 (18%)	1 (1%)	42 (63%)	1 (1%)	2 (3%)	0	5 (7%)	3 (4%)	66
PRP.stimul‖of	6 (8%)	0	42 (61%)	1 (1%)	8 (11%)	1 (1%)	9 (13%)	1 (1%)	68
VBS.inity	9 (17%)	4 (7%)	31 (59%)	1 (1%)	2 (3%)	0	3 (5%)	2 (3%)	52
VBS.elicit	6 (10%)	4 (7%)	33 (58%)	3 (5%)	2 (3%)	0	5 (8%)	3 (5%)	56

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
PRP.modul‖of	7 (7%)	1 (1%)	55 (**58%**)	2 (2%)	12 (12%)	0	15 (**15%**)	2 (2%)	94
VBS.medy	12 (10%)	8 (6%)	67 (**57%**)	3 (2%)	5 (4%)	0	15 (12%)	6 (5%)	116
VBS.diminish	9 (**15%**)	1 (1%)	31 (**54%**)	1 (1%)	8 (14%)	0	6 (10%)	1 (1%)	57
PRP.factor‖in	3 (4%)	26 (**36%**)	38 (**52%**)	1 (1%)	2 (2%)	0	2 (2%)	0	72
VBS.influenc	32 (**17%**)	3 (1%)	95 (**51%**)	7 (3%)	20 (10%)	0	24 (12%)	5 (2%)	186
VBS.augment	11 (**15%**)	2 (2%)	35 (**50%**)	2 (2%)	8 (11%)	0	9 (13%)	2 (2%)	69
VBS.enhanc	47 (**15%**)	8 (2%)	151 (**50%**)	5 (1%)	52 (**17%**)	0	25 (8%)	11 (3%)	299
VBS.control	4 (3%)	23 (**22%**)	52 (**50%**)	2 (1%)	5 (4%)	1 (0%)	12 (11%)	4 (3%)	103
VBS.suppress	44 (**15%**)	20 (7%)	139 (**50%**)	2 (0%)	23 (8%)	0	43 (**15%**)	5 (1%)	276
VBS.attenu	25 (**17%**)	21 (14%)	70 (**48%**)	4 (2%)	6 (4%)	0	12 (8%)	6 (4%)	144
VBS.modul	37 (14%)	6 (2%)	118 (**47%**)	6 (2%)	31 (12%)	0	46 (**18%**)	6 (2%)	250
VBS.restor	16 (**20%**)	0	36 (**46%**)	0	15 (**19%**)	0	7 (9%)	3 (3%)	77
VBS.disrupt	8 (10%)	0	37 (**46%**)	25 (**31%**)	1 (1%)	0	6 (7%)	3 (3%)	80
VBS.abolish	23 (**22%**)	5 (4%)	48 (**46%**)	3 (2%)	8 (7%)	0	12 (11%)	5 (4%)	104
VBS.alter	43 (**21%**)	4 (2%)	90 (**45%**)	12 (6%)	20 (10%)	0	24 (12%)	3 (1%)	196
VBS.stimul	45 (11%)	3 (0%)	174 (**44%**)	8 (2%)	78 (**19%**)	1 (0%)	70 (**17%**)	14 (3%)	393
PRP.precursor‖for	27 (**45%**)	0	26 (**44%**)	0	0	0	5 (8%)	1 (1%)	59
VBS.affect	87 (**18%**)	9 (1%)	199 (**43%**)	29 (6%)	47 (10%)	1 (0%)	73 (**15%**)	15 (3%)	460
VBS.improve	16 (12%)	33 (**25%**)	57 (**43%**)	2 (1%)	8 (6%)	0	13 (9%)	3 (2%)	132
VBS.prevent	39 (**17%**)	52 (**23%**)	92 (**42%**)	7 (3%)	6 (2%)	1 (0%)	14 (6%)	6 (2%)	217
VBS.revers	22 (**18%**)	18 (**15%**)	49 (**41%**)	2 (1%)	6 (5%)	1 (0%)	17 (14%)	3 (2%)	118

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.block	75 (23%)	4 (1%)	129 (39%)	14 (4%)	28 (8%)	0	62 (19%)	13 (4%)	325
VBS.regul	33 (11%)	1 (0%)	113 (39%)	9 (3%)	36 (12%)	0	86 (30%)	8 (2%)	286
VBS.inhibit	140 (13%)	24 (2%)	415 (38%)	24 (2%)	190 (17%)	1 (0%)	231 (21%)	42 (3%)	1067
PRP.marker‖of	8 (14%)	18 (31%)	22 (38%)	3 (5%)	1 (1%)	0	4 (7%)	1 (1%)	57
VBS.chang	20 (25%)	2 (2%)	29 (37%)	4 (5%)	8 (10%)	0	13 (16%)	1 (1%)	77
VBS.maintain	17 (32%)	1 (1%)	19 (35%)	8 (15%)	4 (7%)	0	4 (7%)	0	53
PRP.determin‖of	27 (36%)	7 (9%)	26 (35%)	2 (2%)	4 (5%)	0	7 (9%)	1 (1%)	74
VBS.modify	57 (37%)	5 (3%)	54 (35%)	6 (3%)	7 (4%)	0	21 (13%)	4 (2%)	154
VBS.induce	63 (10%)	118 (19%)	211 (34%)	16 (2%)	38 (6%)	2 (0%)	141 (22%)	26 (4%)	615
PRP.inducer‖of	13 (9%)	10 (7%)	45 (33%)	2 (1%)	14 (10%)	0	43 (32%)	7 (5%)	134
VBS.caus	35 (14%)	81 (33%)	79 (32%)	10 (4%)	3 (1%)	0	28 (11%)	5 (2%)	241
PRP.inhibitor‖of	77 (10%)	10 (1%)	224 (31%)	21 (2%)	96 (13%)	2 (0%)	268 (37%)	20 (2%)	718
VBS.antagon	25 (32%)	5 (6%)	24 (31%)	2 (2%)	6 (7%)	0	13 (16%)	2 (2%)	77
PRP.product‖of	170 (53%)	0	98 (30%)	2 (0%)	11 (3%)	0	35 (11%)	1 (0%)	317
VBS.increas	117 (24%)	7 (1%)	148 (30%)	17 (3%)	56 (11%)	1 (0%)	118 (24%)	17 (3%)	481
VBS.reduc	156 (32%)	33 (6%)	144 (30%)	10 (2%)	42 (8%)	0	72 (15%)	18 (3%)	475
PRP.factor‖for	4 (3%)	56 (55%)	30 (29%)	1 (0%)	2 (1%)	1 (0%)	4 (3%)	3 (2%)	101
VBS.decreas	96 (27%)	18 (5%)	102 (29%)	7 (2%)	42 (12%)	0	72 (20%)	13 (3%)	350
VBS.exhibit	22 (29%)	1 (1%)	21 (28%)	2 (2%)	7 (9%)	0	19 (25%)	2 (2%)	74
VBS.down-regulate	7 (10%)	0	18 (28%)	0	7 (10%)	0	31 (48%)	1 (1%)	64
VBS.show	35 (34%)	5 (4%)	28 (27%)	3 (2%)	6 (5%)	0	20 (19%)	4 (3%)	101

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.requir	27 (**19%**)	5 (3%)	38 (**27%**)	12 (8%)	19 (13%)	3 (2%)	27 (**19%**)	8 (5%)	139
PRP.substrat‖for	10 (6%)	1 (0%)	44 (**27%**)	3 (1%)	11 (6%)	0	92 (**56%**)	1 (0%)	162
VBS.produc	80 (**40%**)	30 (**15%**)	48 (**24%**)	4 (2%)	3 (1%)	1 (0%)	28 (14%)	5 (2%)	199
PRP.activat‖of	8 (7%)	0	25 (**23%**)	5 (4%)	15 (14%)	0	51 (**48%**)	2 (1%)	106
VBS.up-regulate	6 (8%)	0	16 (**23%**)	1 (1%)	8 (11%)	0	38 (**55%**)	0	69
NP2.inhibitor	86 (**15%**)	0	131 (**23%**)	27 (4%)	48 (8%)	0	254 (**44%**)	23 (4%)	569
NP2.product	80 (**55%**)	0	31 (**21%**)	4 (2%)	3 (2%)	0	24 (**16%**)	2 (1%)	144
NP2.blocker	30 (**21%**)	0	29 (**20%**)	4 (2%)	24 (**17%**)	0	50 (**35%**)	4 (2%)	141
VBS.mimic	48 (**57%**)	4 (4%)	16 (**19%**)	4 (4%)	2 (2%)	0	7 (8%)	2 (2%)	83
VBS.lower	31 (**46%**)	1 (1%)	11 (**16%**)	1 (1%)	7 (10%)	0	12 (**18%**)	3 (4%)	66
VBS.activat	57 (14%)	1 (0%)	63 (**16%**)	17 (4%)	46 (12%)	2 (0%)	186 (**48%**)	9 (2%)	381
PRP.agent‖for	11 (**18%**)	34 (**58%**)	9 (**15%**)	2 (3%)	0	0	1 (1%)	1 (1%)	58
PRP.compon‖of	131 (**50%**)	3 (1%)	39 (**15%**)	51 (**19%**)	1 (0%)	1 (0%)	23 (8%)	10 (3%)	259
Relations that have a preference for arguments of type: GO_CELL_COMPONENT									
VBS.enter	11 (**15%**)	2 (2%)	10 (14%)	31 (**44%**)	0	0	12 (**17%**)	3 (4%)	69
VBS.reach	19 (**32%**)	3 (5%)	1 (1%)	20 (**34%**)	0	0	14 (**24%**)	1 (1%)	58
VBS.disrupt	8 (10%)	0	37 (**46%**)	25 (**31%**)	1 (1%)	0	6 (7%)	3 (3%)	80
VBS.stabil	42 (**43%**)	3 (3%)	5 (5%)	23 (**23%**)	7 (7%)	0	15 (**15%**)	1 (1%)	96
PRP.constitu‖of	54 (**53%**)	0	5 (4%)	20 (**19%**)	1 (0%)	0	14 (13%)	7 (6%)	101
PRP.compon‖of	131 (**50%**)	3 (1%)	39 (**15%**)	51 (**19%**)	1 (0%)	1 (0%)	23 (8%)	10 (3%)	259
VBS.target	19 (**24%**)	1 (1%)	9 (11%)	15 (**19%**)	7 (9%)	0	21 (**27%**)	5 (6%)	77

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
NP2.constitu	41 (**70%**)	0	0	11 (**18%**)	0	0	4 (6%)	2 (3%)	58
NP2.compon	84 (**66%**)	0	3 (2%)	23 (**18%**)	4 (3%)	0	10 (7%)	3 (2%)	127
VBS.protect	43 (**39%**)	4 (3%)	10 (9%)	17 (**15%**)	3 (2%)	4 (3%)	23 (**20%**)	6 (5%)	110
VBS.maintain	17 (**32%**)	1 (1%)	19 (**35%**)	8 (**15%**)	4 (7%)	0	4 (7%)	0	53
Relations that have a preference for arguments of type: GO_MOL_FUNCT									
NP2.protein	20 (**20%**)	1 (1%)	7 (7%)	10 (10%)	20 (**20%**)	2 (2%)	36 (**37%**)	1 (1%)	97
VBS.stimul	45 (11%)	3 (0%)	174 (**44%**)	8 (2%)	78 (**19%**)	1 (0%)	70 (**17%**)	14 (3%)	393
VBS.restor	16 (**20%**)	0	36 (**46%**)	0	15 (**19%**)	0	7 (9%)	3 (3%)	77
VBS.inhibit	140 (13%)	24 (2%)	415 (**38%**)	24 (2%)	190 (**17%**)	1 (0%)	231 (**21%**)	42 (3%)	1067
VBS.enhanc	47 (**15%**)	8 (2%)	151 (**50%**)	5 (1%)	52 (**17%**)	0	25 (8%)	11 (3%)	299
NP2.blocker	30 (**21%**)	0	29 (**20%**)	4 (2%)	24 (**17%**)	0	50 (**35%**)	4 (2%)	141
VBS.inactiv	17 (**25%**)	0	7 (10%)	3 (4%)	11 (**16%**)	0	27 (**39%**)	3 (4%)	68
Relations that have a preference for arguments of type: PROTEIN									
PRP.substrat‖of	9 (12%)	0	8 (10%)	1 (1%)	5 (6%)	0	47 (**64%**)	3 (4%)	73
PRP.ligand‖for	15 (**22%**)	0	2 (2%)	2 (2%)	5 (7%)	1 (1%)	41 (**61%**)	1 (1%)	67
PRP.substrat‖for	10 (6%)	1 (0%)	44 (**27%**)	3 (1%)	11 (6%)	0	92 (**56%**)	1 (0%)	162
$\overbrace{\text{VBS.transport}}$	10 (**18%**)	0	2 (3%)	4 (7%)	5 (9%)	0	30 (**55%**)	3 (5%)	54
VBS.up-regulate	6 (8%)	0	16 (**23%**)	1 (1%)	8 (11%)	0	38 (**55%**)	0	69
$\overbrace{\text{VBS.convert}}$	23 (**26%**)	0	6 (6%)	4 (4%)	7 (8%)	1 (1%)	45 (**51%**)	1 (1%)	87
$\overbrace{\text{VBS.hydrolyz}}$	13 (**19%**)	0	0	6 (9%)	8 (12%)	0	34 (**51%**)	5 (7%)	66
NP2.activat	14 (**18%**)	0	5 (6%)	7 (9%)	11 (14%)	0	39 (**51%**)	0	76

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.activat	57 (14%)	1 (0%)	63 (**16%**)	17 (4%)	46 (12%)	2 (0%)	186 (**48%**)	9 (2%)	381
VBS.down-regulate	7 (10%)	0	18 (**28%**)	0	7 (10%)	0	31 (**48%**)	1 (1%)	64
PRP.activat‖of	8 (7%)	0	25 (**23%**)	5 (4%)	15 (14%)	0	51 (**48%**)	2 (1%)	106
NP2.agonist	68 (**38%**)	0	1 (0%)	6 (3%)	19 (10%)	0	82 (**46%**)	0	176
NP2.inhibitor	86 (**15%**)	0	131 (**23%**)	27 (4%)	48 (8%)	0	254 (**44%**)	23 (4%)	569
$\overbrace{\text{VBS.bind}}$	79 (**38%**)	0	2 (0%)	14 (6%)	11 (5%)	2 (0%)	90 (**43%**)	9 (4%)	207
NP2.antagonist	110 (**46%**)	0	5 (2%)	2 (0%)	19 (8%)	0	100 (**42%**)	1 (0%)	237
PRP.antagonist‖of	27 (**33%**)	1 (1%)	5 (6%)	2 (2%)	11 (13%)	0	34 (**41%**)	1 (1%)	81
NP2.substrat	71 (**49%**)	0	3 (2%)	3 (2%)	6 (4%)	0	59 (**40%**)	2 (1%)	144
VBS.inactiv	17 (**25%**)	0	7 (10%)	3 (4%)	11 (**16%**)	0	27 (**39%**)	3 (4%)	68
$\overbrace{\text{VBS.show}}$	18 (**33%**)	0	3 (5%)	5 (9%)	1 (1%)	3 (5%)	21 (**39%**)	2 (3%)	53
$\overbrace{\text{VBS.rece}}$	27 (**33%**)	7 (8%)	3 (3%)	5 (6%)	0	7 (8%)	31 (**38%**)	0	80
NP2.ligand	52 (**43%**)	0	0	8 (6%)	9 (7%)	0	45 (**37%**)	6 (5%)	120
PRP.inhibitor‖of	77 (10%)	10 (1%)	224 (**31%**)	21 (2%)	96 (13%)	2 (0%)	268 (**37%**)	20 (2%)	718
NP2.protein	20 (**20%**)	1 (1%)	7 (7%)	10 (10%)	20 (**20%**)	2 (2%)	36 (**37%**)	1 (1%)	97
$\overbrace{\text{VBS.have}}$	71 (**41%**)	4 (2%)	3 (1%)	15 (8%)	2 (1%)	7 (4%)	62 (**35%**)	9 (5%)	173
NP2.blocker	30 (**21%**)	0	29 (**20%**)	4 (2%)	24 (**17%**)	0	50 (**35%**)	4 (2%)	141
VBS.bind	58 (**49%**)	0	3 (2%)	7 (5%)	3 (2%)	0	38 (**32%**)	8 (6%)	117
PRP.inducer‖of	13 (9%)	10 (7%)	45 (**33%**)	2 (1%)	14 (10%)	0	43 (**32%**)	7 (5%)	134
$\overbrace{\text{VBS.produc}}$	60 (**37%**)	6 (3%)	20 (12%)	10 (6%)	7 (4%)	4 (2%)	48 (**30%**)	4 (2%)	159
VBS.regul	33 (11%)	1 (0%)	113 (**39%**)	9 (3%)	36 (12%)	0	86 (**30%**)	8 (2%)	286

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.remov	31 (**41%**)	1 (1%)	7 (9%)	3 (4%)	6 (8%)	0	22 (**29%**)	4 (5%)	74
VBS.provid	22 (**40%**)	0	8 (14%)	8 (14%)	1 (1%)	0	16 (**29%**)	0	55
VBS.us	68 (**40%**)	7 (4%)	19 (11%)	9 (5%)	5 (2%)	4 (2%)	47 (**27%**)	10 (5%)	169
VBS.target	19 (**24%**)	1 (1%)	9 (11%)	15 (**19%**)	7 (9%)	0	21 (**27%**)	5 (6%)	77
VBS.lack	22 (**39%**)	2 (3%)	2 (3%)	8 (14%)	0	2 (3%)	15 (**26%**)	5 (8%)	56
VBS.contain	234 (**51%**)	9 (1%)	8 (1%)	55 (12%)	10 (2%)	3 (0%)	121 (**26%**)	18 (3%)	458
VBS.have	148 (**54%**)	6 (2%)	21 (7%)	11 (4%)	11 (4%)	0	71 (**26%**)	4 (1%)	272
VBS.exhibit	22 (**29%**)	1 (1%)	21 (**28%**)	2 (2%)	7 (9%)	0	19 (**25%**)	2 (2%)	74
VBS.us	143 (**60%**)	2 (0%)	11 (4%)	8 (3%)	6 (2%)	3 (1%)	59 (**24%**)	6 (2%)	238
VBS.increas	117 (**24%**)	7 (1%)	148 (**30%**)	17 (3%)	56 (11%)	1 (0%)	118 (**24%**)	17 (3%)	481
VBS.reach	19 (**32%**)	3 (5%)	1 (1%)	20 (**34%**)	0	0	14 (**24%**)	1 (1%)	58
VBS.gener	37 (**44%**)	1 (1%)	12 (14%)	5 (6%)	3 (3%)	1 (1%)	20 (**24%**)	4 (4%)	83
NP2.receptor	48 (**64%**)	0	0	5 (6%)	3 (4%)	0	18 (**24%**)	1 (1%)	75
NP2.activity	25 (**48%**)	0	5 (9%)	2 (3%)	6 (11%)	0	12 (**23%**)	2 (3%)	52
VBS.induce	63 (10%)	118 (**19%**)	211 (**34%**)	16 (2%)	38 (6%)	2 (0%)	141 (**22%**)	26 (4%)	615
VBS.inhibit	140 (13%)	24 (2%)	415 (**38%**)	24 (2%)	190 (**17%**)	1 (0%)	231 (**21%**)	42 (3%)	1067
VBS.protect	43 (**39%**)	4 (3%)	10 (9%)	17 (**15%**)	3 (2%)	4 (3%)	23 (**20%**)	6 (5%)	110
VBS.decreas	96 (**27%**)	18 (5%)	102 (**29%**)	7 (2%)	42 (12%)	0	72 (**20%**)	13 (3%)	350
VBS.show	35 (**34%**)	5 (4%)	28 (**27%**)	3 (2%)	6 (5%)	0	20 (**19%**)	4 (3%)	101
NP2.residu	43 (**70%**)	0	0	2 (3%)	4 (6%)	0	12 (**19%**)	0	61
VBS.requir	27 (**19%**)	5 (3%)	38 (**27%**)	12 (8%)	19 (13%)	3 (2%)	27 (**19%**)	8 (5%)	139

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
PRP.level‖of	42 (**73%**)	0	2 (3%)	0	1 (1%)	0	11 (**19%**)	1 (1%)	57
VBS.block	75 (**23%**)	4 (1%)	129 (**39%**)	14 (4%)	28 (8%)	0	62 (**19%**)	13 (4%)	325
VBS.give	46 (**71%**)	0	4 (6%)	1 (1%)	0	0	12 (**18%**)	1 (1%)	64
VBS.releas	54 (**72%**)	0	1 (1%)	3 (4%)	1 (1%)	0	14 (**18%**)	2 (2%)	75
VBS.modul	37 (14%)	6 (2%)	118 (**47%**)	6 (2%)	31 (12%)	0	46 (**18%**)	6 (2%)	250
VBS.lower	31 (**46%**)	1 (1%)	11 (**16%**)	1 (1%)	7 (10%)	0	12 (**18%**)	3 (4%)	66
VBS.stimul	45 (11%)	3 (0%)	174 (**44%**)	8 (2%)	78 (**19%**)	1 (0%)	70 (**17%**)	14 (3%)	393
PRP.regul‖of	12 (6%)	1 (0%)	132 (**66%**)	5 (2%)	11 (5%)	0	35 (**17%**)	3 (1%)	199
VBS.enter	11 (**15%**)	2 (2%)	10 (14%)	31 (**44%**)	0	0	12 (**17%**)	3 (4%)	69
$\overline{\text{VBS.reduc}}$	111 (**68%**)	1 (0%)	3 (1%)	7 (4%)	6 (3%)	2 (1%)	28 (**17%**)	4 (2%)	162
$\overline{\text{VBS.releas}}$	61 (**52%**)	1 (0%)	12 (10%)	11 (9%)	5 (4%)	1 (0%)	20 (**17%**)	5 (4%)	116
$\overline{\text{VBS.affect}}$	39 (**67%**)	0	5 (8%)	1 (1%)	1 (1%)	1 (1%)	10 (**17%**)	1 (1%)	58
VBS.chang	20 (**25%**)	2 (2%)	29 (**37%**)	4 (5%)	8 (10%)	0	13 (**16%**)	1 (1%)	77
VBS.antagon	25 (**32%**)	5 (6%)	24 (**31%**)	2 (2%)	6 (7%)	0	13 (**16%**)	2 (2%)	77
NP2.product	80 (**55%**)	0	31 (**21%**)	4 (2%)	3 (2%)	0	24 (**16%**)	2 (1%)	144
VBS.constitut	46 (**67%**)	0	1 (1%)	6 (8%)	0	0	11 (**16%**)	4 (5%)	68
VBS.repres	105 (**72%**)	1 (0%)	3 (2%)	5 (3%)	5 (3%)	0	23 (**15%**)	2 (1%)	144
PRP.modul‖of	7 (7%)	1 (1%)	55 (**58%**)	2 (2%)	12 (12%)	0	15 (**15%**)	2 (2%)	94
$\overline{\text{VBS.form}}$	64 (**72%**)	0	3 (3%)	4 (4%)	1 (1%)	0	14 (**15%**)	2 (2%)	88
VBS.affect	87 (**18%**)	9 (1%)	199 (**43%**)	29 (6%)	47 (10%)	1 (0%)	73 (**15%**)	15 (3%)	460
VBS.compris	30 (**58%**)	0	1 (1%)	7 (13%)	0	0	8 (**15%**)	5 (9%)	51

Relationship	CHEBI	D	GO_BP	GO_CC	GO_MF	TAXON	PROT	OTHER	Total
VBS.stabil	42 (**43%**)	3 (3%)	5 (5%)	23 (**23%**)	7 (7%)	0	15 (**15%**)	1 (1%)	96
VBS.suppress	44 (**15%**)	20 (7%)	139 (**50%**)	2 (0%)	23 (8%)	0	43 (**15%**)	5 (1%)	276
VBS.reduc	156 (**32%**)	33 (6%)	144 (**30%**)	10 (2%)	42 (8%)	0	72 (**15%**)	18 (3%)	475
HYPONYM	3247 (**72%**)	110 (2%)	176 (3%)	84 (1%)	107 (2%)	9 (0%)	678 (**15%**)	73 (1%)	4484

Table 33: Frequencies of different semantic types. Abbreviations are as described in Table 6. The first column shows terms describing relationships between entities. Each row shows the different types of arguments `Y` found for the triple `{[Chemical X] [term] [Y]}`. Terms are sorted by those that show the strongest preference for each semantic type, by counts of distinct values for `Y`. Only terms with a total of more than 50 different arguments are shown.

B

SOFTWARE USED

Here I list the software used, to provide an opportunity to give versions and references/links, and to clarify which tools were employed for which sections of the project.

B.1 PROGRAMMING LANGUAGES

B.1.1 *Bash*

Bash is a scripting language that in this project has only been used for the simplest purposes — those of executing programmes and redirecting the output to files, other programmes, or both, so as to specify a pipeline.
GNU bash, version 3.2.25(1)-release was used. Documentation at `http://www.gnu.org/s/bash/`.

B.1.2 *Perl*

Perl is a dynamically-typed platform-independent interpreted multi-paradigm programming language. It was used for processing of XML, scripting, and for integrating input and output to other executables. In particular, the text processing done in chapter 4 for NER was done almost entirely in Perl. Perl was used for its terseness and ease of development (for this programmer), its specialisation for string-processing (regular expressions being one of its basic data types, for example), and for the fact that it interacts well with a *nix system.
Versions 5.8 and 5.10 have been used. Documentation at `http://perldoc.perl.org`.

B.1.3 *Java*

Java is an object-oriented platform independent language that was used here for its speed relative to Perl, for its relative ease of programming, and for the rich choice of libraries and IDEs available.

Java version 6 was used. Documentation at `http://www.java.com`.
Netbeans 6.7.1 was used. Documentation at `http://netbeans.org`.

B.1.4 *XQuery*

XQuery is a declarative language that can be used to search large XML datasets for complex patterns. It is used in an equivalent way to that in which SQL is used for searching relational databases. Annotated samples of XQuery can be found in Appendix C.

B.1.5 *R*

R is a powerful language much used for statistical work, especially in the biological sciences. In this project it was used only for graph generation.

B.2 LIBRARIES

B.2.1 *Perl Libraries*

B.2.1.1 *XML::Twig*

XML::Twig is a library that can parse extremely large XML documents in a standards-compliant way. It can perform efficiently by not reading the entire document into memory at any one time, but processing "twigs" from the "tree" one-by-one. For our purposes, a twig can be defined to correspond to a sentence, and abstract, or a chemical, as necessary for the task in hand.
Version 3.26 was used. Documentation is at `http://www.xmltwig.com/xmltwig`.

B.2.1.2 *Inline::Java*

This module allows Java objects and functions to be wrapped so they can be treated as Perl objects and functions. This enabled the use of tested Java code for the clients to in-house server-based tools within Perl scripts rather than necessitating re-implementation.
Version 0.52 was used. Documentation is at `http://search.cpan.org/~patl/Inline-Java-0.52/`.

B.2.1.3 *Text::English*

This module implements a variant of Porter's stemming algorithm, and is used to perform stemming for the NER step, and some normalisation of terms following relation extraction.
Version 0.01 was used. Documentation is at `http://search.cpan.org/perldoc?Text::English`.

B.2.2 *Java Libraries*

B.2.2.1 *Saxon*

Saxon (`http://saxon.sourceforge.net`) is an XSLT and XQuery library for Java that can operate on XML files that have not been pre-processed. It was used in this project to apply XQuery patterns to the XML output of Enju.

B.2.2.2 *BioJava*

BioJava (`http://www.biojava.org`) is a wide-ranging Java library of functions likely to be of use to bioinformaticians and biologists. In this project its OBO-parsing and ontology-navigating sections were used whenever ontologies needed to be manipulated within Java.

B.3 MACHINE LEARNING TOOLS

B.3.1 *SVM^{hmm}*

SVM^{hmm} (`http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html`) is an implementation of a support vector machine for sequence tagging, in this project used as the basis of the NER system. V3.03 was used.

B.3.2 *Weka*

Weka (*Waikato Environment for Knowledge Analysis*) is a machine learning toolkit that allows a wide variety of common algorithms for classification, clustering, and association to be brought to bear on a data set. It also contains a simple graphical user interface. Weka 3.6 was used. Documentation is at `http://www.cs.waikato.ac.nz/ml/weka/`.

B.3.3 *Apriori*

Apriori is an association rule discovery algorithm. The implementation by Christian Borgelt[19] (`http://www.borgelt.net/apriori.html`), version 5.72, was used.

B.4 MISCELLANEOUS

B.4.1 *Enju*

Enju is a syntactic parser for English, providing POS tagging, phrase structure, and predicate-argument information for sentences. It is provided with a kernel trained on the GENIA dataset for biomedical applications. Enju 2.2 was used.

B.4.2 *monq*

Monq is a software package used within the Text-Mining group at the European Bioinformatics Institute (EBI) to build simple annotation servers that can be chained together into pipelines. These were used to supply the sentencising elements of the NER system (see Chapter 4).

B.4.3 *OSCAR3*

OSCAR3 (`http://www-pmr.ch.cam.ac.uk/wiki/Oscar3`) is a widely-used chemical NER system. In this project, the tokenizer module from OSCAR3 was used in construction of an NER system (see Chapter 4).

B.4.4 *LaTeX*

This thesis is typeset with LaTeX 2_ε, making extensive use of the ClassicThesis package: `http://www.miede.de`. The typeface used is Palatino. The syntactic trees were produced with Qtree (`http://www.ling.upenn.edu/advice/latex/qtree`).

B.4.5 *SQLite*

SQLite (`http://www.sqlite.org`) is a relational database management system that allows the database to be stored as a single file and accessed by programs written in a large number of programming languages. It has the advantage of not requiring a separate server process, and for small databases or where high performance is not a requirement it enables rapid development. It was used in this project as a Java library for storing normalised synonyms for entities from the ontologies, and as a Perl library in the annotation interface.

B.4.6 *Image credits*

The frontispiece illustration contains an image of the face and brain from Bourgery & Jacob's *Anatomie de l'Homme* (1831), an image of a Babbage "difference machine" from Charles Babbage's *Passages from the Life of a Philosopher* (1864) and a 1910 illustration of distillation by alembic.

Figure 1 is a facsimile of p.64 of the 1668 edition of John Wilkins' *An essay towards a real character and a philosophical language*.

Figure 4 is based on `https://commons.wikimedia.org/wiki/File:Cube_2x2x2.svg`, which is released under the GNU Free Documentation License.

Figure 7 is from an 1894 natural history engraving.

All other images were drawn or otherwise generated by the author, using the GIMP, Inkscape, and Winsor & Newton "Indian Red" ink.

C

XQUERY SAMPLES

This chapter contains samples of the XQuery used in implementing the LSPs for extracting relations. A few functions from the FunctX XQuery Functions website* were used; these have a `functx:` namespace.

```
declare function local:head( $x as node()? ) as node()?
{
    if ($x/@base)
    then $x
    else
        if (fn:matches($x/@sem_head, "c"))
        then
            local:head($x/cons[@id = $x/@sem_head])
        else
            local:head($x/tok[@id = $x/@sem_head])
} ;


declare function local:pretty ( $cons as node()? ) as xs:string
{
     fn:string-join( fn:data($cons//tok) , ' ')
} ;

declare function local:is_chem ( $x as node()? ) as xs:boolean
{
   every $t in $x//tok satisfies ($t[@entity_list])
   or
  (some $c in $x//cons satisfies ($c/@id = $x/@sem_head and local:is_chem($c)))

 or (
        $x/@xcat="NX-COOD" and $x/@cat="NX"
        and (
            some $c in $x/cons[COOD]/cons satisfies (local:is_chem($c))
        )
    )

} ;
```

Listing 5: General XQuery library functions

```
for $abstract in doc(base-uri())//MedlineCitation,
   $s in $abstract//sentence,
   $x in $s//cons[ local:is_chem(.) ],
        $type in ("copular", "appositive"),
        $direction in ("forwards", "backwards"),
   $tok in $s//tok[
         (
            ($type = "copular" and @aux="copular")
               or
            ($type = "appositive" and @pred="app_arg12")
         )
         and
         (($direction = "forwards" and @arg1 = $x/@id) or ($direction = "backwards
              " and @arg2 = $x/@id))
```

* http://www.xqueryfunctions.com/

```
      ],
   $desc in $s//cons[
      (@id = $tok/@arg1 or @id = $tok/@arg2) and @id ne $x/@id and
      (
         ( matches(@cat, "^N")  and (  .//tok[1]/@cat = "D" or
                                                  index-of(( "one", "two", "
                                                       three","four","five","-
                                                       NUMBER-"  ), zero-or-one
                                                       ((.//tok[1]/@base)[1]) )
                                                  )
               )
   ],

   $head in local:head($desc),
   $phrase in $desc//cons[
      some $t in .//tok satisfies $t/@id = $head/@id
   ] ,
   $chem in $x//cons[
      (. = $x)
      or

      (  (every $t in .//tok satisfies ($t[@entity_list]))
         and
         (not(every $t in ..//tok satisfies ($t[@entity_list]))  or (../@xcat="NX-
              COOD"))
      )
   ]
return

<BE>{$abstract/PMID}<x>{local:pretty($chem)}</x><y>{local:pretty($phrase)}</y>
<tok>{data($tok)}</tok>
<type>{$type}</type>
<dir>{$direction}</dir>
<S>{$s/@id}{local:pretty($s)}</S>{functx:distinct-deep($abstract//ASB_abbreviation
     )}
</BE>
```

Listing 6: XQuery to detect hypernyms

```
  ,,
for $abstract in doc(base-uri())//MedlineCitation,
        $s in $abstract//sentence,
    $x in $s//cons[ local:is_chem(.) ]
        ,
        $type in ("verb"),
        $direction in ("forwards", "backwards"),
        $tok in $s//tok[

                        @pred="verb_arg12"
                        and
                        (($direction = "forwards" and @arg1 = $x/@id) or (
                             $direction = "backwards" and @arg2 = $x/@id))
                ]
        ,
        $desc in $s//cons[
                (@id = $tok/@arg1 or @id = $tok/@arg2) and @id ne $x/@id and
                (
                        ( matches(@cat, "^N")  and (  .//tok[1]/@cat = "D" or
                                                index-of(( "one", "two", "
                                                     three","four","five","-
                                                     NUMBER-"  ), zero-or-one
                                                     ((.//tok[1]/@base)[1]) )
                                                )
                )
                or
                        matches(@cat, "^Axx")
                )


        ],

        $head in local:head($desc),
        $phrase in ($desc,$desc//cons)[
                some $t in .//tok satisfies $t/@id = $head/@id
        ] ,
        $chem in ($x,$x//cons)[
                (. = $x)
                or

                (  (every $t in .//tok satisfies ($t[@entity_list]))
                        and
                        (not(every $t in ..//tok satisfies ($t[@entity_list]))  or
                              (../@xcat="NX-COOD"))
                )
       ]

return

<BE>{$abstract/PMID}<x>{local:pretty($chem)}</x><y>{local:pretty($phrase)}</y>
<tok>{$tok/@base}{$tok/@pos}{data($tok)}</tok>
<type>{$type}</type>
<dir>{$direction}</dir>
<S>{$s/@id}{local:pretty($s)}</S>{functx:distinct-deep($abstract//ASB_abbreviation
     )}
</BE>
```

Listing 7: XQuery to detect transitive verb relations

```
  ,,
declare function local:is_chem ( $x as node()? ) as xs:boolean {
   every $t in $x//tok satisfies ($t[@entity_list])
or
  (some $c in $x//cons satisfies ($c/@id = $x/@sem_head and local:is_chem($c)))
or (
        $x/@xcat="NX-COOD" and $x/@cat="NX"
        and (
            some $c in $x/cons[COOD]/cons satisfies (local:is_chem($c))
        )
```

```
   )
};
```

Listing 8: XQuery to detect chemicals

```
  ,,
declare function local:head( $x as node()? ) as node()? {
    if ($x/@base)
    then $x
    else
        if (fn:matches($x/@sem_head, "c"))
        then
            local:head($x/cons[@id = $x/@sem_head])
        else
            local:head($x/tok[@id = $x/@sem_head])
} ;
```

Listing 9: XQuery to identify the semantic head of a phrase by recursive descent

D

CAFFEINE - A CASE STUDY

Table 34 contains all the properties (including hypernyms) extracted for the molecule Caffeine (`CHEBI:27732`) with entities longer than three characters. It is included partly for comparison (indirectly) with the properties extracted in Giles and Wren(2008)[45], and partly as a general example of typical data for a reasonably commonly-attested chemical entity. The properties are sorted by descending frequency of occurrence. The properties have been assessed by the author and are colour-coded as True ; False ; Partially accurate, or useful but needing clarification .

Freq	Relation	Entity	
57	`is_a`	inhibitor	CHEBI:35222
55	`is_a`	antagonist	CHEBI:48706
44	`is_a`	drug	CHEBI:23888
21	NP2.antagonist	adenosine receptor	PR:000001439
15	`is_a`	agonist	CHEBI:48705
12	`is_a`	methylxanthine	CHEBI:25348
11	NP2.substance	psychotropic drug	CHEBI:35471
10	`is_a`	central nervous system stimulant	CHEBI:35337
8	NP2.antagonist	adenosine	CHEBI:16335

Freq	Relation	Entity	
7	is_a	psychotropic drug	CHEBI:35471
7	is_a	negative regulation of kinase activity	GO:0033673
Negative regulat*or* of kinase activity			
7	is_a	alkaloid	CHEBI:22315
6	VBS.have	protein IMPACT	PR:000009019
This is a mis-resolution of "impact" in the phrase *Caffeine has an impact [on . . .]*			
6	NP2.stimulant	central nervous system drug	CHEBI:35470
This is a result of *Central nervous system drug* being abbreviated to *Central nervous system* for resolution			
6	NP2.drug	psychotropic drug	CHEBI:35471
5	PRP.antagonist‖of	adenosine receptor	PR:000001439
5	is_a	probe	CHEBI:50406
4	VBS.inhibit	phosphorylation	GO:0016310
4	VBS.abolish	phosphorylation	GO:0016310
4	NP2.inhibitor	phosphoric diester hydrolase activity	GO:0008081
4	NP2.ingredient	psychotropic drug	CHEBI:35471
4	NP2.agonist	ryanodine-sensitive calcium-release channel activity	GO:0005219
4	is_a	metabolite	CHEBI:25212
4	is_a	adjuvant	CHEBI:60809
3	VBS.release	calcium(2+)	CHEBI:29108

Freq	Relation	Entity	
3	VBS.activate	ryanodine-sensitive calcium-release channel activity	GO:0005219
3	PRP.antagonist‖at	adenosine receptor	PR:000001439
3	NP2.enhancer	3′,5′-cyclic AMP	CHEBI:17489
3	NP2.derivative	xanthine	CHEBI:15318
3	NP2.derivative	methylxanthine	CHEBI:25348
3	NP2.alkaloid	purine	CHEBI:35584
3	is_a	xanthine	CHEBI:15318
3	is_a	purine alkaloid	CHEBI:26385
2	VBS.suppress	kinase activity	GO:0016301
2	VBS.stimulate	central nervous system drug	CHEBI:35470
This is a result of *Central nervous system drug* being abbreviated to *Central nervous system* for resolution			
2	VBS.override	DNA damage checkpoint	GO:0000077
2	VBS.inhibit	transport	GO:0006810
2	VBS.inhibit	metabolic process	GO:0008152
2	VBS.inhibit	biosynthetic process	GO:0009058
2	VBS.increase	metabolic process	GO:0008152
2	VBS.have	role	CHEBI:50906
2	VBS.block	phosphorylation	GO:0016310
2	VBS.affect	developmental process	GO:0032502
2	PRP.inhibitor‖of	phosphoric diester hydrolase activity	GO:0008081

Freq	Relation	Entity	
2	PRP.effects‖of	paracetamol	CHEBI:46195
Mis-parsing of lists as appositive structure: *The effects of paracetamol, caffeine and [. . .]*			
2	PRP.activator‖of	ryanodine-sensitive calcium-release channel activity	GO:0005219
2	NP2.order	tumor necrosis factor receptor superfamily member 11A	PR:000001954
Mis-resolution of "*rank*" in "*Rank Order*"			
2	NP2.modulator	ryanodine-sensitive calcium-release channel activity	GO:0005219
2	NP2.drug	probe	CHEBI:50406
2	NP2.combination	drug	CHEBI:23888
Caffeine itself is not a drug combination, but it is a drug that is used in combination (in the cases picked up here, with ephedrine)			
2	NP2.antagonist	purinergic receptor activity	GO:0035586
2	NP2.analogue	xanthine	CHEBI:15318
2	NP2.a	adenosine	CHEBI:16335
Mis-parsing of sentences like "*caffeine , the selective adenosine A (2A) antagonist*"			
2	`is_a`	ryanodine receptor modulator	CHEBI:38809
2	`is_a`	phosphodiesterase inhibitor	CHEBI:50218
2	`is_a`	molecule	CHEBI:25367
2	`is_a`	dextromethorphan	CHEBI:4470

Freq	Relation	Entity	
List mis-parsed as appositive structure			
2	is_a	bronchodilator agent	CHEBI:35523
2	is_a	adenosine A2A receptor antagonist	CHEBI:53121
1	VBS.ward	Parkinson's disease	DOID:14330
Sentence here hedged "*caffeine may ward off Parkinson's disease*"			
1	VBS.unaffected	transport	GO:0006810
For which read "*did not affect*". Context: "*The serosal transport was unaffected by caffeine*"			
1	VBS.trigger	cell death	GO:0008219
1	VBS.trigger	calcium(2+)	CHEBI:29108
Triggers Ca^{2+} *release*			
1	VBS.treat	apnea of prematurity	DOID:11163
1	VBS.suppress	sleep	GO:0030431
1	VBS.suppress	phosphorylation	GO:0016310
1	VBS.suppress	growth	GO:0040007
1	VBS.suppress	cell growth	GO:0016049
1	VBS.suppress	binding	GO:0005488
1	VBS.support	role	CHEBI:50906
1	VBS.stimulate	transient receptor potential cation channel TRPV1 isoform 1	PR:000001067
Misresolution of "*alpha-*"			
1	VBS.stimulate	transcription, DNA-dependent	GO:0006351

Freq	Relation	Entity	
1	VBS.stimulate	transcriptional regulator modE	PR:000023270
1	VBS.stimulate	signaling threshold-regulating transmembrane adapter 1	PR:000014894
Misresolution of *"sites"*			
1	VBS.stimulate	caffeine	CHEBI:27732
1	VBS.stimulate	binding	GO:0005488
1	VBS.shorten	cell	GO:0005623
1	VBS.shift	pyraclofos	CHEBI:38876
Misresolution of *"voltage"*			
1	VBS.remove	mannose permease IIC component	PR:000023162
Misresolution of *"many"*			
1	VBS.relieve	phosphorylation	GO:0016310
1	VBS.release	calcium atom	CHEBI:22984
1	VBS.regulate	nuclear mRNA cis splicing, via spliceosome	GO:0045292
1	VBS.regulate	gene expression	GO:0010467
1	VBS.reduce	poly(hydroxyalkanoate)	CHEBI:53387
Misresolution of *"phases"*			
1	VBS.reduce	phosphorylation	GO:0016310
1	VBS.reduce	inositol 1,4,5 trisphosphate binding	GO:0070679
Negation: *"[. . .] but caffeine did not reduce specific [3H] InsP3 binding to the receptor"*			
1	VBS.reduce	binding	GO:0005488

Freq	Relation	Entity	
1	VBS.reach	cell	GO:0005623
1	VBS.quantify	developmental process	GO:0032502
Misparsing (subject of "*quantifies*" is "*model*" rather than "*caffeine*") and misresolution of "*development*" - context "*We propose a [. . .] model for caffeine that quantifies the development of tolerance to[. . .]*".			
1	VBS.protect	membrane	GO:0016020
1	VBS.protect	cell	GO:0005623
1	VBS.protect	caspase-14	PR:000005054
Misresolution of "*mice*"			
1	VBS.promote	conditioned taste aversion	GO:0001661
1	VBS.produce	syndrome	DOID:225
1	VBS.produce	diuresis	GO:0030146
1	VBS.produce	behavior	GO:0007610
1	VBS.prevent	negative regulation of transcription by glucose	GO:0045014
Negation - context: "*Caffeine substantially decreased glucose consumption and growth but did not increase beta-galactosidase activity and did not prevent glucose repression*"			
1	VBS.orientate	cell	GO:0005623
1	VBS.modulate	tumor necrosis factor production	GO:0032640
1	VBS.modulate	gene expression	GO:0010467
1	VBS.mediate	intracellular signal transduction	GO:0035556
1	VBS.lower	signal_peptide	SO:0000418
1	VBS.inhibit	tracer	CHEBI:35204

Freq	Relation	Entity	
Misparse - Inhibits tracer *incorporation*			
1	VBS.inhibit	signal transduction	GO:0007165
1	VBS.inhibit	serine-protein kinase ATM	PR:000004427
1	VBS.inhibit	positive regulation of NF-kappaB transcription factor activity	GO:0051092
1	VBS.inhibit	necrosis	GO:0008220
1	VBS.inhibit	kinase activity	GO:0016301
1	VBS.inhibit	intestinal absorption	GO:0050892
1	VBS.inhibit	growth	GO:0040007
1	VBS.inhibit	epidermal growth factor receptor binding	GO:0005154
1	VBS.inhibit	cell migration	GO:0016477
1	VBS.inhibit	cell cycle arrest	GO:0007050
1	VBS.inhibit	binding	GO:0005488
1	VBS.inhibit	ATP binding	GO:0005524
1	VBS.inhibit	ATPase activity	GO:0016887
1	VBS.inhibit	adenosine receptor	PR:000001439
1	VBS.influence	protein IMPACT	PR:000009019
Misresolution of "*impact*"			
1	VBS.induce	mitochondrion	GO:0005739
Misparse of "*Contribution of mitochondria to the removal of intracellular Ca^{2+} induced by caffeine*"			
1	VBS.induce	metabolic process	GO:0008152

Freq	Relation	Entity	
1	VBS.induce	apoptotic process	GO:0006915
1	VBS.increase	thymidine kinase	PR:000024045
Misparse of *"[. . .] caffeine significantly increased the thymidine kinase (Tk) mutation frequencies [. . .]"*			
1	VBS.increase	motor activity	GO:0003774
Polysemy — *"motor activity"* is being used in the sense of movement at a whole-organism (rat) level			
1	VBS.increase	gene expression	GO:0010467
1	VBS.increase	brain-derived neurotrophic factor	PR:000004716
1	VBS.increase	binding	GO:0005488
1	VBS.increase	behavior	GO:0007610
1	VBS.exacerbate	developmental process	GO:0032502
1	VBS.evoke	glutaryl-7-aminocephalosporanic-acid acylase activity	GO:0033968
Misresolution of *"[␣Ca"*			
1	VBS.enhance	transcriptional regulator modE	PR:000023270
Misresolution of *"mode"*			
1	VBS.enhance	positive regulation of mitochondrial membrane permeability	GO:0035794
1	VBS.enhance	induction of apoptosis	GO:0006917
1	VBS.enhance	binding	GO:0005488
1	VBS.elicit	glucose intolerance	DOID:10603

Freq	Relation	Entity	
Hedged statement: "*We examined whether or not Caf would elicit a glucose intolerance [. . .]*" (and the result of the study was that there was no such effect).			
1	VBS.elicit	diuresis	GO:0030146
1	VBS.displace	binding	GO:0005488
1	VBS.displace	2,3,7,8-tetrachlorodibenzodioxine	CHEBI:28119
1	VBS.deplete	calcium atom	CHEBI:22984
1	VBS.demonstrate	role	CHEBI:50906
1	VBS.delay	habituation	GO:0046959
1	VBS.decrease	localization	GO:0051179
1	VBS.decrease	binding	GO:0005488
1	VBS.change	signal_peptide	SO:0000418
"*Signal*" here is not the signal peptide			
1	VBS.cause	mental disorder	DOID:0050329
1	VBS.cause	acrosome reaction	GO:0007340
1	VBS.block	signal_peptide	SO:0000418
1	VBS.block	localization	GO:0051179
1	VBS.block	kinase activity	GO:0016301
1	VBS.block	developmental process	GO:0032502
1	VBS.block	adenosine receptor	PR:000001439
1	VBS.augment	reflex	GO:0060004
1	VBS.attenuate	vasodilation	GO:0042311

Freq	Relation	Entity	
1	VBS.antagonize	conjugated linoleic acid	CHEBI:61159
1	VBS.alter	reflex	GO:0060004
1	VBS.alter	gene expression	GO:0010467
1	VBS.alter	conditioned taste aversion	GO:0001661
1	VBS.affect	synaptic transmission	GO:0007268
1	VBS.affect	myofibril	GO:0030016
1	VBS.affect	metabolic process	GO:0008152
1	VBS.affect	growth	GO:0040007
The sentence that this was derived from was negated: "*caffeine in pregnancy doesn't affect the baby's growth*". However there are other sentences such as "*[. . .] caffeine inhibits the growth of hepatocellular carcinoma (HCC) cells*" that support the assertion in general			
1	VBS.affect	biosynthetic process	GO:0009058
1	VBS.affect	binding	GO:0005488
1	VBS.activate	SPARC	PR:000015475
Misresolution of "*ones*"			
1	VBS.activate	signal transduction	GO:0007165
1	VBS.activate	narrow pore, gated channel activity	GO:0022831
1	VBS.activate	caspase-14	PR:000005054
Misresolution of "*mice*"			
1	VBS.abrogate	traversing start control point of mitotic cell cycle	GO:0007089
1	VBS.abrogate	serine/threonine-protein kinase ATR	PR:000004499

Freq	Relation	Entity	
Misparse: caffeine abrogates ATR-mediated delay rather than ATR *"[...] abrogate the ATR- and Chk1-mediated delay in progression through S-phase"*.			
1	VBS.abrogate	catabolic process	GO:0009056
1	PRP.theophylline␣(␣TH␣)‖in	gelsolin isoform 1	PR:000002327
Misresolution of *"plasma"*			
1	PRP.structural␣analogue‖at	extracellular-glycine-gated chloride channel activity	GO:0016934
Misparse: *"caffeine is a structural analogue of strychnine and a competitive antagonist at ionotropic glycine receptors"*			
1	PRP.small␣molecule␣inhibitor‖of	kinase activity	GO:0016301
1	PRP.reversal‖in	chlordiazepoxide hydrochloride	CHEBI:3612
Misresolution of *"balance"*			
1	PRP.relative‖of	theophylline	CHEBI:28177
1	PRP.psychostimulant‖on	cognition	GO:0050890
1	PRP.portion‖of	excretion	GO:0007588
"urinary excretion" is being used to signify the substance rather than the process. Context *"[...] caffeine is a minor portion of urinary excretion."*			
1	PRP.nutritional␣precipitating␣factors‖of	migraine	DOID:6364
1	PRP.non-selective␣antagonists‖of	adenosine	CHEBI:16335
1	PRP.more␣efficient␣releaser‖than	noradrenaline	CHEBI:33569
1	PRP.markers‖of	cytochrome P450 1A2	PR:000006102

Freq	Relation	Entity	
1	PRP.in_vivo_probe‖for	dimethylaniline monooxygenase [N-oxide-forming] 3	PR:000007576
Negation: "*Therefore , benzydamine , but not caffeine , is a potential in vivo probe for human FMO3*"			
1	PRP.intervention‖for	flour treatment agent	CHEBI:64577
Misresolution of "*improving*"			
1	PRP.initial_drug‖for	apnea of prematurity	DOID:11163
1	PRP.inhibitor‖of	signal transduction	GO:0007165
1	PRP.inhibitor‖of	positive regulation of NF-kappaB transcription factor activity	GO:0051092
1	PRP.inhibitor‖of	DNA repair	GO:0006281
1	PRP.inhibitor‖of	calcineurin	CHEBI:53439
Misparse of "*These mutants are also sensitive to hygromycin B, caffeine, and FK506, a specific inhibitor of calcineurin*" — "inhibitor" only applies to FK506			
1	PRP.inhibitor‖of	adenosine receptor	PR:000001439
1	PRP.increase‖in	calcium(2+)	CHEBI:29108
Misparse; an increase in Ca^{2+} is the *response* to caffeine, not caffeine itself			
1	PRP.important_constituents‖of	protein NUT	PR:000011532
Misresolution of "*nuts*"			
1	PRP.galenic_form‖of	caffeine	CHEBI:27732
"*Time release caffeine*" is the galenic form of caffeine.			
1	PRP.effects‖on	behavior	GO:0007610

Freq	Relation	Entity	
Misidentification of hypernymy in "*Caffeine : effects of acute and chronic exposure on the behavior of neonatal rats*"			
1	PRP.effect‖of	drug	CHEBI:23888
Misparse *"[. . .] appeared to be a constant effect of standard anxiety-inducing drugs : caffeine, pentylenetetrazole [. . .]"*			
1	PRP.different␣doses‖of	bleomycin	CHEBI:22907
Misparse of list as appositive structure			
1	PRP.compound‖in	protein BEAN	PR:000004718
Misresolution of "*bean*"			
1	PRP.complex‖with	water	CHEBI:15377
1	PRP.antagonist‖for	adenosine	CHEBI:16335
1	PRP.analogue‖of	strychnine	CHEBI:28973
1	PRP.agonist‖of	ryanodine-sensitive calcium-release channel activity	GO:0005219
1	PRP.activator‖of	signaling	GO:0023052
1	NP2.thyroxine	dexamethasone	CHEBI:41879
Misparse of list as appositive structure			
1	NP2.therapy	adjuvant	CHEBI:60809
1	NP2.teratogen	Homo sapiens	NCBITaxon:9606
Negation: "*However, overwhelming evidence indicates that caffeine is not a human teratogen*"			
1	NP2.stimulus	sensory perception of taste	GO:0050909
1	NP2.stimulant	lipopolysaccharide-induced tumor necrosis factor-alpha factor	PR:000009843

Freq	Relation	Entity	
Misresolution of "*simple*"			
1	NP2.shift	chlordiazepoxide hydrochloride	CHEBI:3612
Misresolution of "*balance*"			
1	NP2.releaser	calcium atom	CHEBI:22984
1	NP2.portion	nuclear receptor subfamily 4 group A member 3	PR:000011410
Misresolution of "*minor*"			
1	NP2.neuron	serotonergic drug	CHEBI:48278
Misparse identifying "*serotonergic neuron*" as in apposition to caffeine			
1	NP2.moclobemide	inhibitor	CHEBI:35222
Misparse identifying "*the inhibitor moclobemide*" in list as in apposition to caffeine			
1	NP2.mobilizer	calcium(2+)	CHEBI:29108
1	NP2.methylxanthine	alkaloid	CHEBI:22315
1	NP2.level	calcium(2+)	CHEBI:29108
Misparse identifying appositive structure in "*on removal of caffeine , the SR Ca(2+) levels partially recovered*"			
1	NP2.intake	calcium atom	CHEBI:22984
Misparse of list as appositive structure			
1	NP2.inhibitor	molecule	CHEBI:25367
1	NP2.inhibitor	DNA repair	GO:0006281
1	NP2.inducer	cytochrome P450 1A2	PR:000006102
1	NP2.h–NUMBER-	inhibitor	CHEBI:35222

Freq	Relation	Entity	
Misparse of list as appositive structure			
1	NP2.gly-leu	peptide	CHEBI:16670
Misparse of list as appositive structure			
1	NP2.ephedrine	magnesium-25 atom	CHEBI:52763
Misparse of list as appositive structure and mis-resolution of "*25 mg ephedrine*".			
1	NP2.derivative	purine	CHEBI:35584
1	NP2.cost	peroxy group	CHEBI:29369
Misparse of list as appositive structure and misresolution of "*O(2)*"			
1	NP2.constituent	psychotropic drug	CHEBI:35471
1	NP2.constituent	food	CHEBI:33290
1	NP2.compound	methylxanthine	CHEBI:25348
1	NP2.chemical	psychotropic drug	CHEBI:35471
1	NP2.blocker-caffeine	beta-adrenergic drug	CHEBI:48540
Misparse of "*[. . .] compared with caffeine alone, the beta-adrenergic blocker-caffeine combination[. . .]*"			
1	NP2.blocker	adenosine receptor	PR:000001439
1	NP2.a-NUMBER-	adenosine receptor	PR:000001439
Misparse of "*A1 and A2A adenosine receptor antagonist*"			
1	NP2.a-NUMBER-	adenosine	CHEBI:16335
Misparse of "*A1, A2A, and A2B adenosine receptor antagonist*"			
1	NP2.analogue	base	CHEBI:22695

Freq	Relation	Entity	
1	NP2.alkaloid	trimethylxanthine	CHEBI:27134
1	NP2.alkaloid	beta-carboline	CHEBI:109895
misparse *"[. . .] caffeine and eudistomin D , a beta-carboline alkaloid[. . .]"*			
1	NP2.adjuvant	analgesic	CHEBI:35480
1	is_a	vasoconstrictor agent	CHEBI:50514
1	is_a	tryptophan	CHEBI:27897
Misparse of list as appositive structure			
1	is_a	theophylline	CHEBI:28177
Misparse of list as appositive structure			
1	is_a	teratogenic agent	CHEBI:50905
Negation			
1	is_a	serine-protein kinase ATM	PR:000004427
Misparse; Caffeine is a serine-protein kinase ATM *inhibitor*			
1	is_a	secondary metabolite	CHEBI:26619
1	is_a	sarcoplasmic reticulum	GO:0016529
1	is_a	reagent	CHEBI:33893
1	is_a	purine	CHEBI:35584
1	is_a	protein IMPACT	PR:000009019
Misresolution of "*impact*"			
1	is_a	protein Dos	PR:000006641

Freq	Relation	Entity	
Misresolution of "*doses*"			
1	is_a	peroxy group	CHEBI:29369
Misparse erroneously identifying appositive structure and misresolution of "*O(2)*"			
1	is_a	nutrient	CHEBI:33284
The sentence is in the form of a question: "*Caffeine: a nutrient, a drug or a drug of abuse*". It is not possible to tell from the English-language abstract what conclusions are drawn			
1	is_a	negative regulation of cyclic-nucleotide phosphodiesterase activity	GO:0051344
1	is_a	mineral	CHEBI:46662
Misparse of list as appositive structure			
1	is_a	indicator	CHEBI:47867
1	is_a	heroin	CHEBI:27808
Misparse of list as appositive structure			
1	is_a	glutamate 5-kinase	PR:000023597
Misresolution of "*probes*"			
1	is_a	food additive	CHEBI:64047
1	is_a	(-)-ephedrine	CHEBI:15407
Misparse of list as appositive structure			
1	is_a	diuretic	CHEBI:35498
1	is_a	dexamethasone	CHEBI:41879
Misparse of list as appositive structure			

Freq	Relation	Entity	
1	is_a	chemical substance	CHEBI:59999
1	is_a	central nervous system drug	CHEBI:35470
1	is_a	catechin	CHEBI:23053
Misparse of list as appositive structure			
1	is_a	antioxidant	CHEBI:22586
1	is_a	alizarin	CHEBI:16866
Misparse of list as appositive structure			
1	is_a	adenosine receptor A1	PR:000001575
Misparse of *"adenosine receptor A1 and A2A receptor antagonist"*			
1	is_a	5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	CHEBI:18406
Misparse of list as appositive structure			
1	is_a	1-(3-chlorophenyl)piperazine	CHEBI:10588
Misparse of list as appositive structure			
1	VBS.wash	signal_peptide	SO:0000418
"signal" misidentified as the subject of "washing" in *"After washing ryanodine and caffeine , the aequorin signal and muscle tone returned to their respective control levels"*			
1	VBS.receive	group	CHEBI:24433
"group" here is not a chemical group but a group of individuals			
1	VBS.inhibit	ruthenium atom	CHEBI:30682
Misparse of *"[...] inhibited both caffeine- and eugenol-induced muscle contractions"*; Misresolution of *"Ruthenium red"*			

Freq	Relation	Entity	
1	$\overleftarrow{\text{VBS.include}}$	nutraceutical	CHEBI:50733
1	$\overleftarrow{\text{VBS.include}}$	inhibitor	CHEBI:35222
1	$\overleftarrow{\text{VBS.consume}}$	SMAD5 antisense gene protein 1	PR:000015255
Misresolution of *"dams"* (in the sense of mothers)			
1	$\overleftarrow{\text{VBS.bind}}$	signaling threshold-regulating transmembrane adapter 1	PR:000014894
Misresolution of *"sites"*			
1	$\overleftarrow{\text{VBS.apply}}$	cell	GO:0005623
"cell" misidentified as the subject of *"applying"* in *"[...] a cell was first initialized to deplete the SR Ca load by applying caffeine."*			
1	$\overleftarrow{\text{VBS.add}}$	cell	GO:0005623
"cell" misidentified as the subject of *"adding"* in *"[...] the cell was stimulated to enter mitosis by adding 10 mM caffeine."*			

Table 34: Information extracted about caffeine

E

SCREENSHOTS FROM CURATION INTERFACE

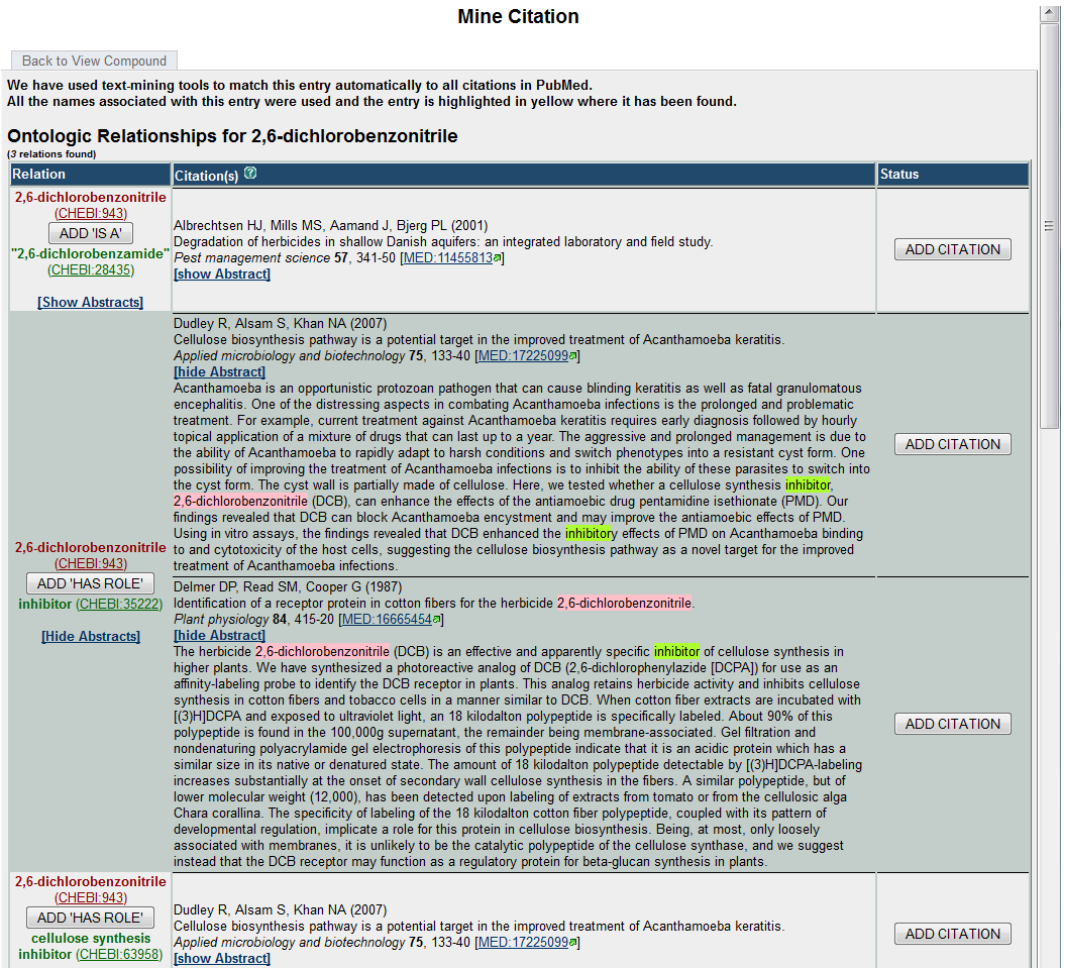


Figure 30: Screenshot of curator interface. The data supplied to the software was from the hypernym extraction system; the software itself was adapted for this purpose by Adriano Dekker.

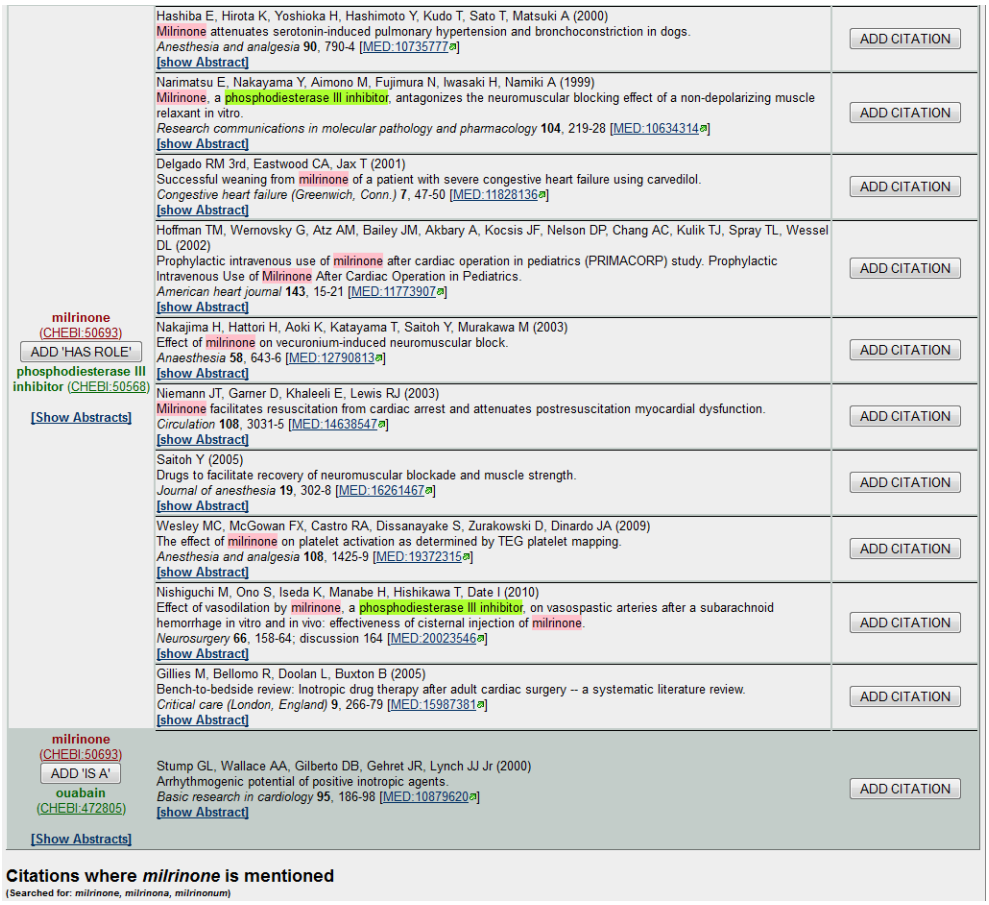


Figure 31: Another screenshot of curator interface. The data supplied to the software was from the hypernym extraction system; the software itself was adapted for this purpose by Adriano Dekker.

BIBLIOGRAPHY

[1] Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In *In Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94, 2000.

[2] Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martinez. Enriching very large ontologies using the WWW, October 2000.

[3] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.*, 22(2):207–216, June 1993. ISSN 0163-5808. doi: 10.1145/170036.170072.

[4] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Advances in knowledge discovery and data mining. chapter Fast discovery of association rules, pages 307–328. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6.

[5] Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, and Xinglong Wang. Assisted curation: does text mining really help. In *In The Pacific Symposium on Biocomputing (PSB*, 2008.

[6] Enrique Alfonseca and Suresh Manandhar. Improving an Ontology Refinement Method with Hyponymy Patterns, 2002.

[7] Enrique Alfonseca and Suresh Manandhar. An Unsupervised Method for General Named Entity Recognition And Automated Concept Discovery. In *In: Proceedings of the 1 st International Conference on General WordNet*, 2002.

[8] Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. Event extraction for systems biology by text mining the literature. *Trends in biotechnology*, 28(7):381–390, July 2010. ISSN 1879-3096. doi: 10.1016/j.tibtech.2010.04.005.

[9] Chinatsu Aone and Mila R. Santacruz. REES: a large-scale relation and event extraction system. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 76–83, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/974147.974158.

[10] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001. ISSN 1531-605X.

[11] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556.

[12] Michael Ashburner, Christopher J. Mungall, and Suzanna E. Lewis. Ontologies for biologists: a community model for the annotation of genomic data. *Cold Spring Harbor symposia on quantitative biology*, 68: 227–235, 2003. ISSN 0091-7451.

[13] Michael Bada and Lawrence Hunter. Enrichment of OBO ontologies. *Journal of biomedical informatics*, 40(3):300–315, June 2007. ISSN 1532-0480. doi: 10.1016/j.jbi.2006.07.003.

[14] Edward H Bendix. *Componential analysis of general vocabulary: the semantic structure of a set of verbs in English, Hindi, and Japanese*. Number v. 32 in International journal of American linguistics. Indiana University, 1966.

[15] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, pages 194–201, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. doi: 10.3115/974557.974586.

[16] Stephan Bloehdorn, Roberto Basili, Marco Cammisa, and Alessandro Moschitti. Semantic Kernels for Text Classification Based on Topological Measures of Feature Similarity. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, volume 0, pages 808–812, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2701-9. doi: 10.1109/icdm.2006.141.

[17] Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, September 2006. ISSN 1477-4054. doi: 10.1093/bib/bbl027.

[18] Olivier Bodenreider, Marc Aubry, and Anita Burgun. Non-lexical approaches to identifying associative relations in the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 91–102, 2005. ISSN 1793-5091.

[19] Christian Borgelt. Efficient Implementations of Apriori and Eclat. In *Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL). CEUR Workshop Proceedings 90*, 2003.

[20] Gosse Bouma and Geert Kloosterman. Mining syntactically annotated corpora with XQuery. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 17–24, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[21] Ted Briscoe, Caroline Gasperin, Ian Lewin, and Andreas Vlachos. Bootstrapping an interactive information extraction system for flybase curation. In Michael Ashburner, Ulf Leser, and Dietrich Rebholz-Schuhmann, editors, *Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives*, number 08131 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[22] Anita Burgun and Olivier Bodenreider. An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. 2005.

[23] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 120–126, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. doi: 10.3115/1034678.1034705.

[24] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996. ISSN 0891-2017. doi: 10.3115/997939.997983.

[25] Scott Cederberg and Dominic Widdows. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 111–118, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119176.1119191.

[26] S. Le Cessie and J. C. Van Houwelingen. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201, 1992. ISSN 00359254. doi: 10.2307/2347628.

[27] Don Chamberlin. XQuery: a query language for XML. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, SIGMOD '03, page 682, New York, NY, USA, 2003. ACM. ISBN 1-58113-634-X. doi: 10.1145/872757.872877.

[28] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990. ISSN 0891-2017.

[29] Philipp Cimiano and Johanna Völker. Text2Onto. In Andrés Montoyo, Rafael Muńoz, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238. Springer Berlin Heidelberg, 2005. doi: 10.1007/11428817_21.

[30] Nigel Collier, Chikashi Nobata, and Jun I. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 201–207, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. doi: 10.3115/990820.990850.

[31] Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation*, 3(2-3):281–332, December 2005. ISSN 1570-7075. doi: 10.1007/s11168-006-6327-9.

[32] Ann Copestake, Peter Corbett, Peter Murray-Rust, Advaith Siddharthan, Simone Teufel, and Ben Waldron. An architecture for language processing for scientific texts. In *In Proceedings of the 4th UK E-Science All Hands Meeting*, 2006.

[33] Peter Corbett and Ann Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(Suppl 11):S4+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-s11-s4.

[34] Peter Corbett and Peter Murray-Rust. High-Throughput Identification of Chemistry in Life Science Texts Computational Life Sciences II. volume 4216 of *Lecture Notes in Computer Science*, chapter 11, pages

107–118. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-45767-1. doi: 10.1007/11875741_11.

[35] Peter Corbett, Colin Batchelor, and Simone Teufel. Annotation of chemical named entities. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 57–64, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[36] Peter Corbett, Colin Batchelor, and Ann Copestake. *Pyridines, pyridine and pyridine rings: disambiguating chemical named entities*. Marrakech, Morocco, 2008.

[37] Francisco M. Couto, Mário J. Silva, and Pedro M. Coutinho. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, April 2007. ISSN 0169023X. doi: 10.1016/j.datak.2006.05.003.

[38] David A. Cruse. *Lexical Semantics (Cambridge Textbooks in Linguistics)*. Cambridge University Press, September 1986. ISBN 0521276438.

[39] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. In *Journal of the American Society for Information Science*, pages 391–407, 1990.

[40] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(Database issue):D344–D350, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm791.

[41] Doug Downey, Oren Etzioni, and Stephen Soderland. A probabilistic model of redundancy in information extraction. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, pages 1034–1041, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[42] Mariano Fernández-López. Overview of methodologies for building ontologies. 1999.

[43] Blaž Fortuna, Dunja Mladenič, and Marko Grobelnik. Semi-automatic Construction of Topic Ontologies Semantics, Web and Mining. volume

4289 of *Lecture Notes in Computer Science*, chapter 8, pages 121–131. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-47697-9. doi: 10.1007/11908678_8.

[44] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 233–237, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. ISBN 1-55860-272-0. doi: 10.3115/1075527.1075579.

[45] Cory Giles and Jonathan Wren. Large-scale directional relationship extraction and resolution. *BMC Bioinformatics*, 9(Suppl 9):S11+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-s9-s11.

[46] Julien Gobeill, Emilie Pasche, Dina Vishnyakova, and Patrick Ruch. Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. *Database*, 2013:bat041+, January 2013. ISSN 1758-0463. doi: 10.1093/database/bat041.

[47] Harsha Gurulingappa, Corinna Kolářik, Martin Hofmann-Apitius, and Juliane Fluck. Concept-Based Semi-Automatic Classification of Drugs. *J. Chem. Inf. Model.*, 49(8):1986–1992, August 2009. ISSN 1549-9596. doi: 10.1021/ci9000844.

[48] Harsha Gurulingappa, Abdul M. Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, October 2012. ISSN 15320464. doi: 10.1016/j.jbi.2012.04.008.

[49] Thierry Hamon and Adeline Nazarenko. Detection of synonymy links between terms: experiment and results. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins Publishing Company, 2001. ISBN 978 90 272 9816 4.

[50] Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain Natural Language Processing – IJCNLP 2005. volume 3651 of *Lecture Notes in Computer Science*, chapter 18, pages 199–210. Springer Berlin

/ Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-29172-5. doi: 10.1007/11562214_18.

[51] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992154.

[52] Robert Hoehndorf, Anika Oellrich, Michel Dumontier, Janet Kelso, Dietrich R. Schuhmann, and Heinrich Herre. Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics*, 11(1): 441+, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-441.

[53] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. An Overview of Event Extraction from Text. October 2011.

[54] Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: what's beyond PubMed? *Molecular cell*, 21(5):589–594, March 2006. ISSN 1097-2765. doi: 10.1016/j.molcel.2006.02.012.

[55] Mario Jarmasz. Roget's Thesaurus as a Lexical Resource for Natural Language Processing, March 2012.

[56] Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, October 2009. ISSN 0885-6125. doi: 10.1007/s10994-009-5108-8.

[57] Nikiforos Karamanis, Ruth Seal, Ian Lewin, Peter McQuilton, Andreas Vlachos, Caroline Gasperin, Rachel Drysdale, and Ted Briscoe. Natural Language Processing in aid of FlyBase curators. *BMC Bioinformatics*, 9(1):193+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-193.

[58] Martin Kavalec and Vojtěch Svátek. V.: A Study on Automated Relation Labelling in Ontology Learning. In *Ontology Learning from Text: Methods, Evaluation and Applications. IOS*, pages 44–58, 2005.

[59] Jin D. Kim, Tomoko Ohta, and Jun'ichi Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1): 10+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-10.

[60] Corinna Kolářik, Martin Hofmann-Apitius, Marc Zimmermann, and Juliane Fluck. Identification of new drug classification terms in textual resources. *Bioinformatics*, 23(13):i264–i272, July 2007. ISSN 1460-2059. doi: 10.1093/bioinformatics/btm196.

[61] Anna Korhonen, Ilona Silins, Lin Sun, and Ulla Stenius. The first step in the development of Text Mining technology for Cancer Risk Assessment: identifying and organizing scientific evidence in risk assessment literature. *BMC bioinformatics*, 10(1):303+, September 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-303.

[62] Carl Linnaeus. *Systema naturæ, sive, Regna tria naturæsystematice proposita per classes, ordines, genera, & species.* Apud Theodorum Haak, Joannis Wilhelmi de Groot, 1735. doi: 10.5962/bhl.title.877.

[63] Kaihong Liu, William R. Hogan, and Rebecca S. Crowley. Natural Language Processing methods and systems for biomedical ontology learning. *Journal of biomedical informatics*, 44(1):163–179, February 2011. ISSN 1532-0480. doi: 10.1016/j.jbi.2010.07.006.

[64] John Lyons. *Semantics*. Cambridge University Press, November 1977. ISBN 0521291860.

[65] Mitchell P. Marcus, Mary A. Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017.

[66] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748.

[67] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, December 1990. ISSN 1477-4577. doi: 10.1093/ijl/3.4.235.

[68] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6.

[69] Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages

121–130, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6.

[70] Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. Corpus-Oriented Grammar Development for Acquiring a Head-Driven Phrase Structure Grammar from the Penn Treebank Natural Language Processing – IJCNLP 2004. volume 3248 of *Lecture Notes in Computer Science*, chapter 72, pages 684–693. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-24475-2. doi: 10.1007/978-3-540-30211-7_72.

[71] Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1017–1024, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220303.

[72] Emmanuel Morin and Christian Jacquemin. Automatic acquisition and expansion of hypernym links. In *Computer and the humanities*, pages 363–396, 2003.

[73] Fleur Mougin, Anita Burgun, and Olivier Bodenreider. Using WordNet to improve the mapping of data elements to UMLS for data sources integration. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 574–578, 2006. ISSN 1942-597X.

[74] Christopher Mungall, Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, January 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-1-r2.

[75] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January 2007. ISSN 0378-4169. doi: 10.1075/li.30.1.03nad.

[76] Prakash M. Nadkarni, Lucila Ohno-Machado, and Wendy W. Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, September 2011. ISSN 1527-974X. doi: 10.1136/amiajnl-2011-000464.

[77] Preslav Nakov, Ariel Schwartz, Brian Wolf, and Marti Hearst. Supporting Annotation Layers for Natural Language Processing. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 65–68, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1225753.1225770.

[78] Darren A. Natale, Cecilia N. Arighi, Winona C. Barker, Judith Blake, Ti-Cheng C. Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, and Cathy H. Wu. Framework for a protein ontology. *BMC bioinformatics*, 8 Suppl 9(Suppl 9):S1+, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-s9-s1.

[79] Mariana Neves and Ulf Leser. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, pages bbs084+, December 2012. ISSN 1477-4054. doi: 10.1093/bib/bbs084.

[80] Philip V. Ogren, K. Bretonnel Cohen, George Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. The compositional structure of Gene Ontology terms. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 214–225, 2004. ISSN 1793-5091.

[81] John Osborne, Jared Flatow, Michelle Holko, Simon Lin, Warren Kibbe, Lihua Zhu, Maria Danila, Gang Feng, and Rex Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10(Suppl 1):S6+, 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-s1-s6.

[82] Martin F. Porter. An algorithm for suffix stripping. *Program*, 3(14): 130–137, October 1980.

[83] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50+, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-50.

[84] Marie-Laure Reinberger and Peter Spyns. Discovering Knowledge in Texts for the Learning of DOGMA-Inspired Ontologies. In *ECAI 2004 Workshop on Ontology Learning and Population*, 2004.

[85] Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *AAAI/IAAI, Vol. 2*, pages 1044–1049, 1996.

[86] Ellen Riloff and Jessica Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. In *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124, 1997.

[87] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3+, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-s3-s3.

[88] Fabio Rinaldi, Gerold Schneider, and Simon Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*, 45(5):851–861, October 2012. ISSN 15320464. doi: 10.1016/j.jbi.2012.04.014.

[89] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 517–528, 2000. ISSN 2335-6936.

[90] Thomas C. Rindflesch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, December 2003. ISSN 15320464. doi: 10.1016/j.jbi.2003.11.003.

[91] Frank Rogers. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116, January 1963. ISSN 0025-7338.

[92] Gerard Salton and Chris Buckley. Term Weighting Approaches in Automatic Text Retrieval. Technical report, Ithaca, NY, USA, 1987.

[93] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–462, 2003. ISSN 2335-6936.

[94] Isabel Segura-Bedmar, Paloma Martínez, and María Segura-Bedmar. Drug name recognition and classification in biomedical texts. *Drug Discovery Today*, 13(17-18):816–823, September 2008. ISSN 13596446. doi: 10.1016/j.drudis.2008.06.001.

[95] Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew L. Tan. Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13*, BioMed '03, pages 49–56, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1118958.1118965.

[96] Sidney Siegel and N. John Castellan. *Nonparametric Statistics for The Behavioral Sciences*. McGraw-Hill Humanities/Social Sciences/Languages, 2 edition, January 1988. ISBN 0070573573.

[97] Frank Smadja. Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19(1):143–177, March 1993. ISSN 0891-2017.

[98] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007. ISSN 1087-0156. doi: 10.1038/nbt1346.

[99] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning Syntactic Patterns for Automatic Hypernym Discovery. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA, 2005.

[100] Mark Stevenson, Yikun Guo, Robert Gaizauskas, and David Martinez. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-s11-s7.

[101] Lin Sun, Anna Korhonen, Ilona Silins, and Ulla Stenius. User-Driven Development of Text Mining Resources for Cancer Risk Assessment. 2009.

[102] Simone Teufel. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization (Center for the Study of Language and Information)*. Center for the Study of Language and Inf, March 2010. ISBN 1575865564.

[103] Anne E. Thessen, Hong Cui, and Dmitry Mozzherin. Applications of Natural Language Processing in Biodiversity Science. *Advances in Bioinformatics*, 2012:1–17, 2012. ISSN 1687-8027. doi: 10.1155/2012/391574.

[104] Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. Construction of an annotated corpus to support biomedical informa-

tion extraction. *BMC Bioinformatics*, 10(1):349+, October 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-349.

[105] Peter D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42536-5.

[106] Kimberly Van Auken, Joshua Jaffery, Juancarlos Chan, Hans M. Muller, and Paul Sternberg. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, 10(1):228+, July 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-228.

[107] Jorge E. Villaverde, Agustín Persson, Daniela Godoy, and Analía Amandi. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. *Expert Systems with Applications*, 36 (7):10288–10294, September 2009. ISSN 09574174. doi: 10.1016/j.eswa.2009.01.048.

[108] Thomas Wächter and Michael Schroeder. Semi-automated ontology generation within OBO-Edit. *Bioinformatics (Oxford, England)*, 26(12): i88–i96, June 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq188.

[109] Yanli Wang, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang, and Stephen H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37 (Web Server issue):W623–W633, July 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp456.

[110] Jonathan J. Webster and Chunyu Kit. Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics - Volume 4*, COLING '92, pages 1106–1110, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992424.992434.

[111] Dominic Widdows and Beate Dorow. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1072228.1072342.

[112] John Wilkins. *An essay towards a real character: and a philosophical language*. Printed for S. Gellibrand, 1668.

[113] Rainer Winnenburg, Thomas Wächter, Conrad Plake, Andreas Doms, and Michael Schroeder. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in Bioinformatics*, 9(6):466–478, November 2008. ISSN 1477-4054. doi: 10.1093/bib/bbn043.

[114] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005. ISBN 0120884070.

[115] Tao Xu, LinFang Du, and Yan Zhou. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9(1):472+, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-472.

[116] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, March 2003. ISSN 1532-4435.