Exploring variation across biomedical subdomains

Tom Lippincott and **Diarmuid Ó Séaghdha** and **Lin Sun** and **Anna Korhonen**
Computer Laboratory
University of Cambridge
United Kingdom
{tl318,do242,ls418,alk23}@cam.ac.uk

Abstract

Previous research has demonstrated the importance of handling differences between domains such as "newswire" and "biomedicine" when porting NLP systems from one domain to another. In this paper we identify the related issue of *subdomain variation*, i.e., differences between subsets of a domain that might be expected to behave homogeneously. Using a large corpus of research articles, we explore how subdomains of biomedicine vary across a variety of linguistic dimensions and discover that there is rich variation. We conclude that an awareness of such variation is necessary when deploying NLP systems for use in single or multiple subdomains.

1 Introduction

One of the most noticeable trends in the past decade of Natural Language Processing (NLP) research has been the deployment of language processing technology to meet the information retrieval and extraction needs of scientists in other disciplines. This meeting of fields has proven mutually beneficial: scientists increasingly rely on automated tools to help them cope with the exponentially expanding body of publications in their field, while NLP researchers have been spurred to address new conceptual problems in theirs. Among the fundamental advances from the NLP perspective has been the realisation that tools which perform well on textual data from one source may fail to do so on another unless they are tailored to the new source in some way. This has led to significant interest in the idea of contrasting *domains* and the concomitant problem of *domain adaptation*, as well as the production of manually annotated domain-specific corpora.[1]

One definition of *domain variation* associates it with differences in the underlying probability distributions from which different sets of data are drawn (Daumé III and Marcu, 2006). The concept also mirrors the notion of variation across thematic subjects and the corpus-linguistic notions of *register* and *genre* (Biber, 1988). In addition to the differences in vocabulary that one would expect to observe, domains can vary in many linguistic variables that affect NLP systems. The scientific domain which has received the most attention (and is the focus of this paper) is the biomedical domain. Notable examples of corpus construction projects for the biomedical domain are PennBioIE (Kulick et al., 2004) and GENIA (Kim et al., 2003). These corpora have been used to develop systems for a range of processing tasks, from entity recognition (Jin et al., 2006) to parsing (Hara et al., 2005) to coreference resolution (Nguyen and Kim, 2008).

An implicit assumption in much previous work on biomedical NLP has been that particular subdomains of biomedical literature – typically molecular biology – can be used as a model of biomedical language in general. For example, GENIA consists of abstracts dealing with a specific set of subjects in molecular biology, while PennBioIE covers abstracts in two specialised domains, cancer genomics and the behaviour of a particular class of enzymes. This assumption of representativeness is understandable because lin-

[1] A workshop dedicated to domain adaptation is collocated with ACL 2010.

guistic annotation is labour-intensive and it may not be worthwhile to produce annotated corpora for multiple subdomains within a single discipline if there is little task-relevant variation across those subdomains. However, such conclusions should not be made before studying the actual degree of difference between the subdomains of interest.

One of the principal goals of this paper is to map how the concept of "biomedical language", often construed as a monolithic entity, is composed of diverse patterns of behaviour at more fine-grained topical levels. Hence we study linguistic variation in a broad biomedical corpus of abstracts and full papers, the PMC Open Access Subset.[2] We select a range of lexical and structural phenomena for quantitative investigation. The results indicate that common subdomains for resource development are not representative of biomedical text in general and furthermore that different linguistic features often partition the subdomains in quite different ways.

2 Related Work

A number of researchers have explored the differences between non-technical and scientific language. Biber and Gray (2010) describe two distinctive syntactic characteristics of academic writing which set it apart from general English. Firstly, in academic writing additional information is most commonly integrated by pre- and post-modification of phrases rather than by the addition of extra clauses. Secondly, academic writing places greater demands on the reader by omitting non-essential information, through the frequent use of passivisation, nominalisation and noun compounding. Biber and Gray also show that these tendencies towards "less elaborate and less explicit" language have become more pronounced in recent history.

We now turn to corpus studies that focus on biomedical writing. Verspoor et al. (2009) use measurements of lexical and structural variation to demonstrate that Open Access and subscription-based journal articles in a specific domain (mouse genomics) are sufficiently similar that research on the former can be taken as representative of the latter. While their primary goal is different from ours and they do not consider variation across multiple domains, they do compare their mouse genomics corpus with small reference corpora drawn from newswire and general biomedical sources. This analysis unsurprisingly finds differences between the domain and newswire corpora across many linguistic dimensions; more interestingly for our purposes, the comparison of domain text to the broader biomedical superdomain shows a more complex picture with similarities in some aspects (e.g., passivisation and negation) and dissimilarities in others (e.g., sentence length, semantic features).

Friedman et al. (2002) document the "sublanguages" associated with two biomedical domains: clinical reports and molecular biology articles. They set out restricted ontologies and frequent co-occurrence templates for the two domains and discuss the similarities and differences between them, but they do not perform any quantitative analysis.

Other researchers have focused on specific phenomena, rather than cataloguing a broad scope of variation. Cohen et al. (2008) carry out a detailed analysis of argument realisation with respect to verbs and nominalisations, using the GENIA and PennBioIE corpora. Nguyen and Kim (2008) compare the behaviour of anaphoric pronouns in newswire and biomedical corpora; they improve the performance of a pronoun resolver by incorporating their observations, thus demonstrating the importance of capturing domain-specific phenomena. Nguyen and Kim's findings are discussed in more detail in Section 5.4 below.

3 Subdomains in the OpenPMC Corpus

The Open Access Subset of PubMed (OpenPMC) is the largest publicly available corpus of full-text articles in the biomedical domain. OpenPMC is comprised of 169,338 articles drawn from 1233 medical journals, totalling approximately 400 million words. The NIH maintains a one-to-many mapping from journals to 122 subject areas (NIH, 2009b). This covers about 400 of the OpenPMC journals, but these account for over 70% of the

[2] http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html

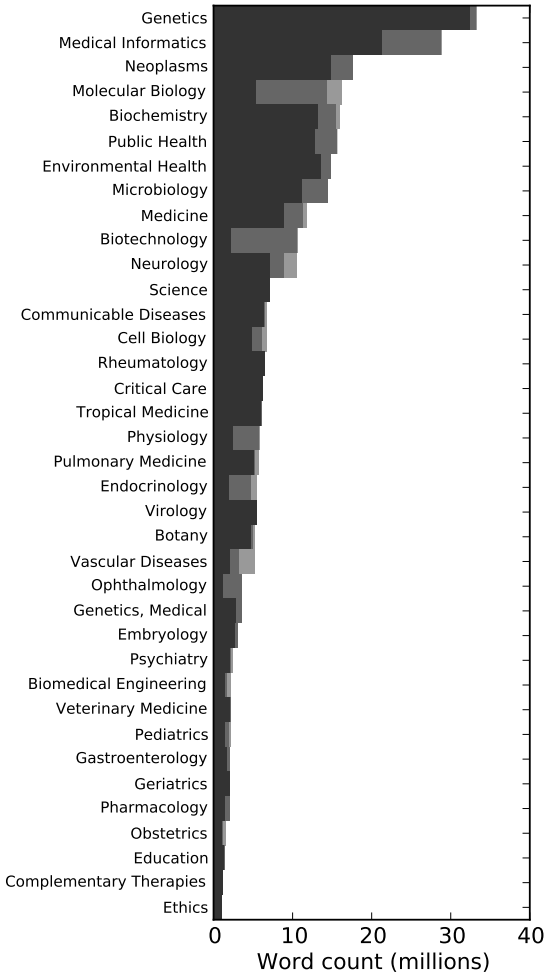


Figure 1: OpenPMC word count by subdomain, dark colouring indicates data assigned single subdomain, each lighter shade indicates an additional overlapping subdomain

database by byte size and word count. Journals are assigned up to five subject areas with the majority assigned one (69%) or two (26%) subjects. In this paper we adopt the OpenPMC subject areas (e.g. "Pulmonary Medicine", "Genetics", "Psychiatry") as the basis for subdomain comparison.

4 Methodology

4.1 Data selection and preprocessing

An important initial question was how to treat data with multiple classifications: we only consider journals assigned a single subdomain, to avoid the added complexity of interactions in data from overlapping subdomains. To ensure sufficient data for comparing a variety of linguistic features, we discard the subdomains with less than one million words meeting the single-subdomain criterion. After review, we also drop the "Biology" subdomain, which appears to function as a catch-all for many loosely related areas. Figure 1 shows the distribution of data across the subjects we use, by word-count, with lighter-coloured areas representing data that is assigned multiple subjects. These subjects provide a convenient starting point for dividing the corpus into subdomains (hereafter, "subdomain" will be used rather than "subject"). We also add a reference subdomain, "Newswire", composed of a 6 million word random sample from the English Gigaword corpus (Graff et al., 2005). The final data set has a total of 39 subdomains.

Articles in the OpenPMC corpus are formatted according to a standard XML tag set (NIH, 2009a). We first convert each article to plain text, ignoring "non-content" elements such as tables and formulas, and split the result into sentences, aggregating the results by subdomain.

4.2 Feature extraction

We investigate subdomain variation in our corpus across a range of lexical, syntactic, sentential and discourse features. The corpus is lemmatised, tagged and parsed using the C&C pipeline (Curran et al., 2007) with the adapted part-of-speech and lexical category tagging models produced by Rimell and Clark (2009) for biomedical parsing.

From this output we count occurrences of noun, verb, adjective and adverb lemmas, part-of-speech (POS) tags, grammatical relations (GRs), chunks, and lexical categories. The lemma features are Zipfian-distributed items from an open class, so we have experimented with filtering low-frequency items at various thresholds to reduce noise and improve processing speed. The other feature sets can be viewed as closed classes, where filtering is unnecessary.

Since verbs are central to the meaning and structure of sentences, we consider their special behavior by constructing features for each verb's distribution over other grammatical properties.

Subdomain	VB	VBG	VBN	VBP	VBZ
Medical Informatics	.35	.29	.06	.09	.21
Cell Biology	.14	.43	.05	.10	.29

Table 1: Distribution over POS tags for verb "restrict", in two subdomains

Several grammatical properties are captured by pairing each verb with its POS (indicating e.g. tense, such as present, past, and present participle). Voice is determined from additional annotation output by the C&C parser. Figure 1 shows the POS-distribution for the verb "restrict", in two subdomains from the corpus. Finally, we record distributions over verb subcategorization frames (SCFs) taken by each verb, and over the GRs it participates in. SCFs were extracted using a system of Preiss et al. (2007).

To facilitate a more robust and interpretable analysis of vocabulary differences, we estimate a "topic model" of the corpus with Latent Dirichlet Analysis (Blei et al., 2003) using the MALLET toolkit.[3] As preprocessing we divide the corpus into articles, removing stopwords and words shorter than 3 characters. The Gibbs sampling procedure is parameterised to induce 100 topics, each giving a coherent cluster of related words learned from the data, and to run for 1000 iterations. We collate the predicted distribution over topics for each article in a subdomain, weighted by article wordcount, to produce a topic distribution for the subdomain.

4.3 Measurements of divergence

Our goal is to illustrate the presence or absence of differences between the feature sets, and to do so we calculated the Jensen-Shannon divergence and the Pearson correlation. Jensen-Shannon divergence is a finite symmetric measurement of the divergence between probability distributions, while Pearson correlation quantifies the linear relationship between two real-valued samples.

The count-features are weighted, for a given subdomain, by the feature's log-likelihood between the subdomain's data and the rest of the corpus. Log-likelihood has been shown to perform well when comparing counts of potentially low-frequency features (Rayson and Garside, 2000) such as found in Zipfian-distributed data. This serves to place more weight in the comparison on items that are distinctive of the subdomain with respect to the entire corpus.

While the count-features are treated as a single distribution for the purposes of JSD, the verbwise-features are composed of many distributions, one for each verb lemma. Our approach is to combine the JSD of the verbs, weighted by the log-likelihood of the verb lemma between the two subdomains in question, and normalize the distances to the interval [0, 1]. Using the lemma's log-likelihood assumes that, when a verb's distribution behaves differently in a subdomain, its frequency changes as well.

We present the results as dendrograms and heat maps. Dendrograms are tree structures that illustrate the results of hierarchical clustering. We perform hierarchical clustering on the inter-subdomain divergences for each set of features. The algorithm begins with each instance (in our case, subdomains) as a singleton cluster, and repeatedly joins the two most similar clusters until all the data is clustered together. The order of these merges is recorded as a tree structure that can be visualized as a dendrogram in which the length of a branch represents the distance between its child nodes. Similarity between clusters is calculated using average distance between all members, known as "average linking".

Heat maps show the pairwise calculation of a metric in a grid of squares, where square x, y is shaded according to the value of $metric(sub_x, sub_y$. For our measurements of JSD, black represents 0 (i.e. identical distributions) and white represents the metric's theoretical maximum of 1. We also inscribe the actual value inside each square. Dendrograms are tree structures that illustrate the hierarchical clustering procedure described above. The dendrograms present all 39 subdomains, while for readibility the heatmaps present 12 subdomains selected for representativeness.

[3] http://mallet.cs.umass.edu

5 Results

Different thresholds for filtering low-frequency terms had little effect on the divergence measures, and served mainly to improve processing time. We therefore report results using a cutoff of 150 occurrences (over the entire 234 million word data set) and log-likelihood weights. The results of Pearson correlation and JSD show similar trends, and due to its specific design for comparing distributions we only report the latter.

5.1 Vocabulary and lexical features

Differences in vocabulary are what first comes to mind when describing subdomains. Word features are fundamental components for systems such as POS taggers and lexicalised parsers; one therefore expects that these systems will be affected by variation in lexical distributions. Figure 2a uses JSD calculated on each subdomain's distribution over 100 LDA-induced topics to compare vocabulary distributions. Subdomains related to molecular biology (Genetics, Molecular Biology) show the smallest divergences, an interesting fact since these are heavily used in building resources for BioNLP. The dendrogram shows a rough division into "public policy", "patient-centric", "applied" and "microscopic" subdomains, with the distance between unrelated subdomains such as Biochemistry and Pediatrics almost as large as their respective differences from Newswire.

We omit figures for variation over noun, verb and adjective lemmas due to space restrictions; in general, these correlate with the variation in LDA topics though there are some differences. Figure 2b shows JSD calculated on distributions over adverb lemmas. Part of the variation is due to characteristic markers of scientific argument ("therefore", "significantly", "statistically"). A more interesting factor is the coining of domain-specific adverbs, an example of the tendency in scientific text to use complex lexical items and premodifiers rather than additional clauses. This also has the effect of moving subdomain-specific objects and processes from verbs and nouns to adverbs. This behavior seems non-continuous, in that subdomains either make heavy, or almost no, use of it: for example, Pediatrics has no subdomain-specific items among the its ten top adverbs by log-likelihood, while Neoplasms has "histologically", "immunohistochemically" and "subcutaneously". These information-dense terms could prove useful for tasks like automatic curation of subdomain vocabularies, where they imply relationships between their components, the items they modify, etc.

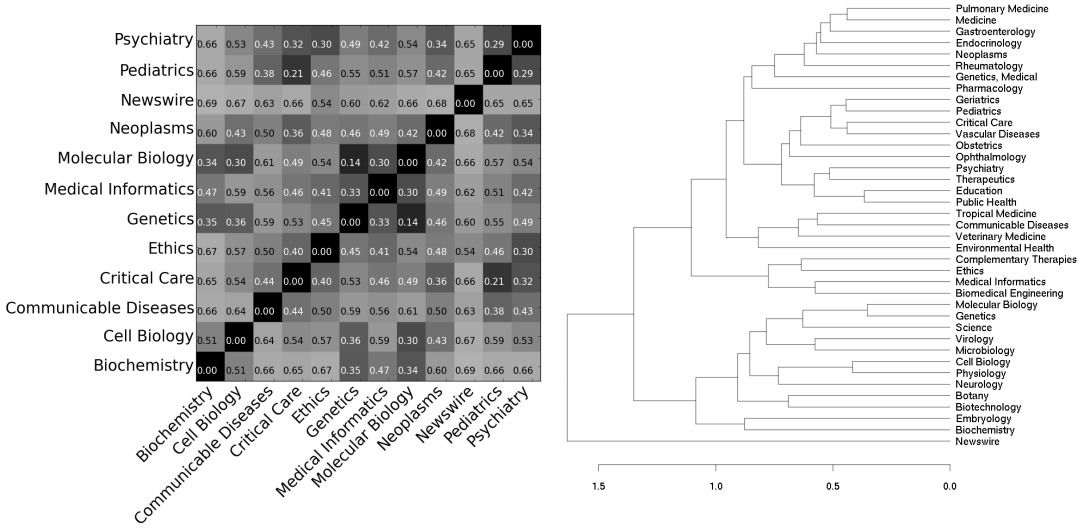
5.2 Verb distributional behavior

Modelling verb behavior is important for both syntactic (Collins, 2003) and semantic (Korhonen et al., 2008) processing, and subdomains are known to conscript verbs into specific roles that change the distributions of their syntactic properties (Roland and Jurafsky, 1998). The four properties we considered verbs' distributions over (SCF, POS, GR and voice) produced similar inter-subdomain JSD values. Figure 2c demonstrates how verbs differ between subdomains with respect to SCFs. For example, while the Pediatrics subdomain uses the verb "govern" in a single SCF among its 12 possibilities, the Genetics subdomain distributes its usage over 7 of them. Two subdomains may both use "restrict" with high frequency (e.g. Molecular Biology and Ethics), but with different frequency distributions over SCFs.

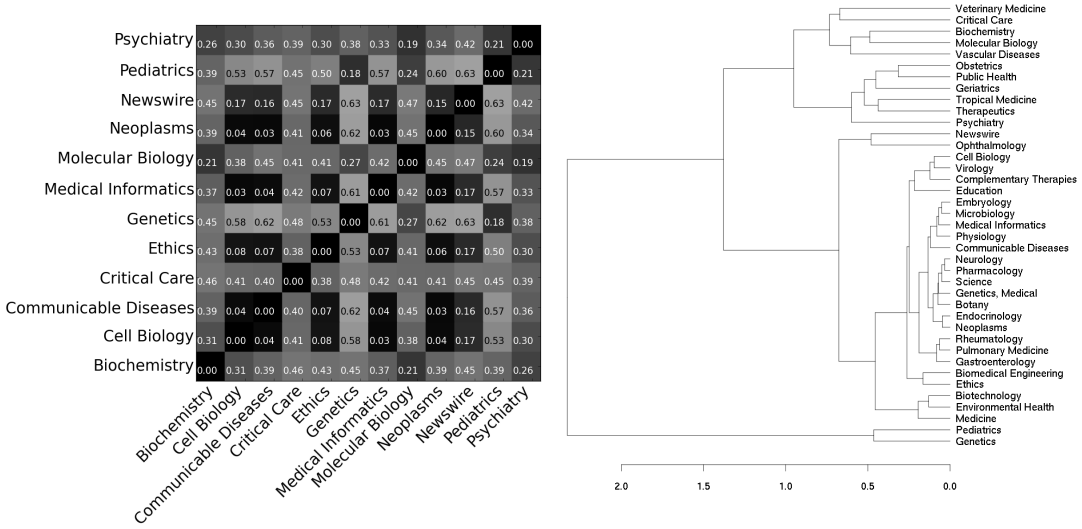
5.3 Syntax

It is difficult to measure syntactic complexity accurately without access to a hand-annotated treebank, but it is well-known that sentence length correlates strongly with processing difficulty (Collins, 1996). The first column of Table 2 gives average sentence lengths (excluding punctuation and "sentences" of fewer than three words) for selected domains. All standard errors are < 0.1. It is clear that all biomedical subdomains typically use longer sentences than newswire, though there is also variation within biomedicine, from an average length of 27 words in Molecular Biology to 24.5 words in Pediatrics.

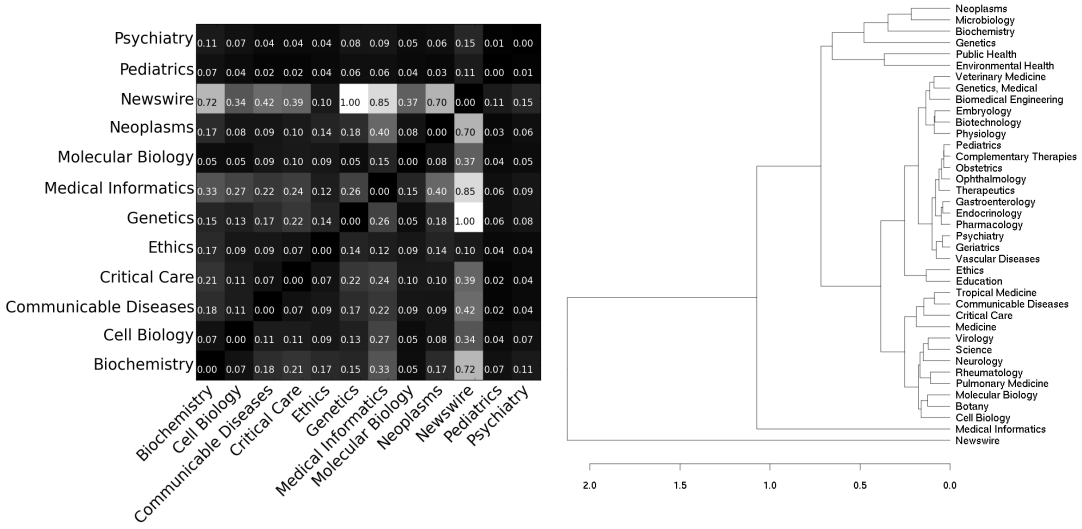
"Packaging" information in complex pre- and/or post-modified noun phrases is a characteristic feature of academic writing (Biber and Gray, 2010). This increases the information density of a sentence but brings with it syntactic and



(a) LDA-induced distribution over topics



(b) Adverb lemma frequencies



(c) Verb distributions over subcategorization frames

Figure 2: Subdomain variation plotted as heat maps and dendrograms

Sentence length		Full NP length		Base nominal length	
Mol. Biology	27.0	Biochemistry	4.03	Biochemistry	1.85
Genetics	26.6	Genetics	3.90	Neoplasms	1.85
Cell Biology	26.3	Critical Care	3.86	Mol. Biology	1.84
Ethics	26.2	Neoplasms	3.85	Genetics	1.83
PMC Average	25.9	PMC Average	3.85	PMC Average	1.80
Biochemistry	25.8	Pediatrics	3.84	Cell Biology	1.80
Neoplasms	25.5	Med. Informatics	3.84	Critical Care	1.80
Psychiatry	25.3	Comm. Diseases	3.81	Med. Informatics	1.78
Critical Care	25.0	Therapeutics	3.80	Comm. Diseases	1.78
Therapeutics	24.9	Mol. Biology	3.79	Therapeutics	1.75
Comm. Diseases	24.9	Psychiatry	3.77	Psychiatry	1.75
Med. Informatics	24.6	Ethics	3.69	Pediatrics	1.73
Pediatrics	24.6	Cell Biology	3.55	Ethics	1.65
Newswire	19.1	Newswire	3.18	Newswire	1.60

Table 2: Average sentence, NP and base nominal lengths across domains

semantic ambiguities. For example, the difficulty of resolving the internal structure of noun-noun compounds and strings of prepositional phrases has been the focus of ongoing research in NLP; these phenomena have also been identified as significant challenges in biomedical language processing (Rosario and Hearst, 2001; Schuman and Bergler, 2006). The second and third columns of Table 2 present average lengths for full noun phrases, defined as every word dominated by a head noun in the grammatical relation graph for a sentence, and for base nominals, defined as nouns plus premodifying adjectives and nouns only. All standard errors are ≤ 0.01. Newswire text uses the simplest noun phrase structures; there is notable variation across PMC domains. Full NP and base nominal lengths do not always correlate; for example, Cell Biology uses relatively long base NPs (nominalisations and multitoken names in particular) but relatively simple full NP structures.

5.4 Coreference

Resolving coreferential terms is a crucial and challenging task when extracting information from texts in any domain. Nguyen and Kim (2008) compare the use of pronouns in the newswire and biomedical domains, using the GENIA corpus as representative of the latter. Among the differences observed between the domains were the absence of any personal pronouns other than third-person neuter pronouns in the GENIA corpus, and a greater proportion of demonstrative pronouns in GENIA than in the ACE or MUC newswire corpora. Corroborating the importance of domain modelling, Nguyen and Kim demonstrate that tailoring a pronoun resolution system to specific properties of the biomedical domain improves performance.

As our corpus is not annotated for coreference we restrict our attention to types that are reliably coreferential: masculine/feminine personal pronouns (*he*, *she* and case variations), neuter personal pronouns (*they*, *it* and variations) and definite NPs with demonstrative determiners such as *this* and *that*. To filter out pleonastic pronouns we used a combination of the C+C parser's pleonasm tag and heuristics based on Lappin and Leass (1994). To filter out the most common class of non-anaphoric demonstrative NPs we simply discarded any matching the pattern *this...paper|study|article*.

Table 3 presents statistics for selected types of coreferential noun phrases in a number of domains. The results generally agree with the findings of Nguyen and Kim (2008): biomedical text is on average 200 times less likely than news text to use gendered pronouns and twice as

Pronouns (neuter, 3rd)		Pronouns (non-neuter, 3rd)		Demonstrative NPs	
Ethics	0.0658	Newswire	0.0591	Genetics	0.0275
Newswire	0.0607	Ethics	0.0037	Med. Informatics	0.0263
Therapeutics	0.0354	Pediatrics	0.0015	Biochemistry	0.0263
Med. Informatics	0.0346	Psychiatry	0.0009	Ethics	0.0260
Psychiatry	0.0342	Comm. Diseases	0.0009	Mol. Biology	0.0251
Pediatrics	0.0308	Therapeutics	0.0005	PMC Average	0.0226
PMC Average	0.0284	PMC Average	0.0005	Cell Biology	0.0210
Genetics	0.0275	Critical Care	0.0004	Comm. Diseases	0.0207
Critical Care	0.0272	Neoplasms	0.0002	Neoplasms	0.0205
Mol. Biology	0.0258	Med. Informatics	0.0002	Psychiatry	0.0201
Biochemistry	0.0251	Genetics	0.0001	Critical Care	0.0201
Neoplasms	0.0227	Mol. Biology	2.5×10^{-5}	Therapeutics	0.0192
Cell Biology	0.0217	Biochemistry	2.0×10^{-5}	Pediatrics	0.0191
Comm. Diseases	0.0213	Cell Biology	1.5×10^{-5}	Newswire	0.0118

Table 3: Frequency of coreferential types (proportion of all NPs) across domains

likely to use anaphoric definite noun phrases. At the domain level, however, there is clear variation within the biomedical corpus. In contrast to Nguyen and Kim's observations about GENIA some domains do make non-negligible use of gendered pronouns, most notably Ethics (usually to refer to other scholars) and domains such as Psychiatry and Pediatrics where studies of actual patients are common. All biomedical domains use demonstrative NPs more frequently than newswire and only one (Ethics) matches newswire for frequent use of neuter 3rd-person pronouns.

6 Conclusion

In this paper we have explored the phenomenon of linguistic variation at a finer-grained level than previous NLP research, focusing on subdomains rather than traditional domains such as "newswire" and "biomedicine". We have identified patterns of variation across dimensions of vocabulary, syntax and discourse that are known to be of importance for NLP applications. While the magnitude of variation between subdomains is unsurprisingly less pronounced than between coarser domains, subdomain variation clearly does exist and should be taken into account when considering the generalisability of systems trained and evaluated on specific subdomains, for example molecular biology.

Future work includes directly evaluating the effect of subdomain variation on practical tasks, investigating further dimensions of variation such as nominalisation usage and learning alternative subdomain taxonomies directly from the corpus text. Ultimately, we expect that a more nuanced understanding of subdomain effects will have tangible benefits for many applications of scientific language processing.

Acknowledgements

This work was supported by EPSRC grant EP/G051070/1, the Royal Society (AK) and a Dorothy Hodgkin Postgraduate Award (LS).

References

Biber, Douglas and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1):2–20.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cohen, K. Bretonnel, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158.

Collins, Michael John. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of ACL-96*, Santa Cruz, CA.

Collins, Michael. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the ACL-07 Demo and Poster Sessions*, Prague, Czech Republic.

Daumé III, Hal and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Friedman, Carol, Pauline Kraa, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Graff, David, Junbo Kong, Ke Chen, and Kazuaki Maeda, 2005. *English Gigaword Corpus, 2nd Edition*. Linguistic Data Consortium.

Hara, Tadayoshi, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proceedings of IJCNLP-05*, Jeju Island, South Korea.

Jin, Yang, Ryan T. McDonald, Kevin Lerman, Mark A. Mandel, Steven Carroll, Mark Y. Liberman, Fernando C. Pereira, Raymond S. Winters, and Peter S. White. 2006. Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics*, 7:492.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):i180–i182.

Korhonen, Anna, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *Proceedings of COLING-08*, Manchester, UK.

Kulick, Seth, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the HLT-NAACL-04 Workshop on Linking Biological Literature, Ontologies and Databases*, Boston, MA.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Nguyen, Ngan L.T. and Jin-Dong Kim. 2008. Exploring domain differences for the design of a pronoun resolution system for biomedical text. In *Proceedings of COLING-08*, Manchester, UK.

NIH. 2009a. Journal publishing tag set. http://dtd.nlm.nih.gov/publishing/.

NIH. 2009b. National library of medicine: Journal subject terms. http://wwwcf.nlm.nih.gov/serials/journals/index.cfm.

Preiss, Judita, E.J. Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL-07*, Prague, Czech Republic.

Rayson, Paul and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the ACL-00 Workshop on Comparing Corpora*, Hong Kong.

Rimell, Laura and Stephen Clark. 2009. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, 42(5):852–865.

Roland, Douglas and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of COLING-ACL-98*, Montreal, Canada.

Rosario, Barbara and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMNLP-01*, Pittsburgh, PA.

Schuman, Jonathan and Sabine Bergler. 2006. Postnominal prepositional phrase attachment in proteomics. In *Proceedings of the HLT-NAACL-06 BioNLP Workshop on Linking Natural Language and Biology*, New York, NY.

Verspoor, Karin, K Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and Open Access scientific journals are similar. *BMC Bioinformatics*, 10:183.