Compasses, Magnets, Water Microscopes
Annotation and Analysis of Terminology in a Diachronic Corpus of Scientific Texts

Anne-Kathrin Schumann, Stefan Fischer
Applied Linguistics, Translation and Interpreting
Universität des Saarlandes, Campus A2.2, 66123 Saarbrücken, Germany
annek_schumann@gmx.de, stefan.fischer@uni-saarland.de

Abstract

The specialised lexicon belongs to the most prominent attributes of specialised writing: *Terms* function as semantically dense encodings of specialised concepts, which, in the absence of terms, would require lengthy explanations and descriptions. In this paper, we argue that terms are the result of diachronic processes on both the semantic and the morpho-syntactic level. Very little is known about these processes. We therefore present a corpus annotation project aiming at revealing how terms are coined and how they evolve to fit their function as semantically and morpho-syntactically dense encodings of specialised knowledge. The scope of this paper is two-fold: Firstly, we outline our methodology for annotating terminology in a diachronic corpus of scientific publications. Moreover, we provide a detailed analysis of our annotation results and suggest methods for improving the accuracy of annotations in a setting as difficult as ours. Secondly, we present results of a pilot study based on the annotated terms. The results suggest that terms in older texts are linguistically relatively simple units that are hard to distinguish from the lexicon of general language. We believe that this supports our hypothesis that terminology undergoes diachronic processes of densification and specialisation.

Keywords: terminology, annotation, diachrony

1. Introduction

Diachrony is an under-researched topic in terminology: Research efforts which are directed towards the acquisition and representation of specialised knowledge normally deal with the state of affairs at a given point in time rather than with knowledge evolution. Terminological dynamics have, at best, been researched only for very short time intervals, and have rarely been researched in a quantitative manner. The question which driving forces are behind terminologisation, specialisation or lexical change in restricted domains has, to our knowledge, not even been touched upon. However, diachronic corpus-linguistic work on specialised texts remains incomplete if due attention is not paid to the development of the lexicon.

This paper describes our efforts in the annotation of terminology in English scientific texts from the Royal Society Corpus (RSC) (Khamis et al., 2015). The corpus is based on the Philosophical Transactions and Proceedings of the Royal Society of London, a resource that has already gained some attention for its role in the construction of a specialised discourse in English (Moessner, 2009; Banks, 2008; Banks, 2009). The texts cover a wide range of domains, including biology, chemistry, physics, geography, and medicine. They were often written by so-called *virtuosi*, or gentlemen scientists, as letters containing descriptions of experiments, observations, etc. Some of the texts are translations of reports originally written in a foreign language or summaries of longer pieces. The time period covered is 1665–1869.

In its current state, the RSC contains texts that have already been passed through a number of preprocessing steps: removal of OCR errors, re-ordering of scrambled pages, identification of text boundaries, duplicate detection, and boilerplate removal. Spelling normalisation and token annotation are underway (Khamis et al., 2015). By annotating terminology in a subset of texts sampled from this corpus, we hope to provide a basis for empirical investigations into the development of terminology over the course of more than 200 years.

2. Related Work

The research effort outlined in this paper is related to two rather distant strands of research, diachronic corpus linguistics, on the one hand, and terminology, especially term annotation and the study of terminological dynamics, on the other hand. The major novelty of our work is the combination of these two strands with the aim of gaining empirical insights into the diachronic development of the specialised lexicon of scientific English.

In the past, *diachronic corpus linguistic work* on scientific prose has shown a greater interest in grammatical phenomena rather than the lexicon. An example is the work of Biber and Gray (2011), who analyse small scientific corpora ranging from the 16^{th} century to the 1990s. An interesting result of their study is that noun phrases in English informational writing have become increasingly elaborate over the last two centuries, making academic discourse a "compressed" genre. However, Biber and Gray (2011) do not relate this change to terminology, i.e. genuinely lexical processes. Instead, they argue that the observed patterns of development reflect a general tendency of natural languages towards tightly integrated grammatical structures.

A similar approach to diachronic processes in language is exploited by Banks (2008), who samples research articles in 20-year intervals from the same source of data that constitutes the RSC, namely the Philosophical Transactions of the Royal Society of London. He then goes on by analysing the development of thematic structure, i.e. linguistic items that fill sentence-initial positions, as well as thematic progression in the texts.

Terminology has contributed to our research from an entirely different angle, namely by developing methods for the annotation of terms in text corpora. Annotation is normally done in the context of term extraction, that is, as an auxiliary task necessary for the creation of gold-standard resources needed for evaluation. Previous research efforts have resulted in well-known resources, such as the GENIA (Kim et al., 2003) and CRAFT (Bada et al., 2012) corpora. Another relevant corpus annotation project is the ACL RD-TEC (Zadeh and Handschuh, 2014), which is based on the ACL ARC (Bird et al., 2008). However, as all these resources are synchronic in nature, they are not suitable for studying the research questions posed by us.

The study of conceptual dynamics over time is, beyond term annotation, another terminology-related strand of research that is relevant to the work presented here. In particular, corpus-based approaches to the history of science have emerged recently, especially due to the growing availability of suitable text corpora. However, early work in this area tended to ignore the fundamental role of terminology in the organisation and communication of knowledge in restricted domains. A well-known example of this general-language approach to specialised knowledge is the study by Hall et al. (2008), who isolate topic clusters by running an LDA (Blei et al., 2003) algorithm on their data set. Gupta and Manning (2011) extend this work by combining LDA with pattern-based information extraction methods.

Recent studies have taken a terminological turn. For example, Mariani et al. (2014) use the TermoStat (Drouin, 2004) term extractor in their analysis of the LREC Anthology to isolate topic identifiers. Schumann and QasemiZadeh (2015) also exploit a term extraction method to identify terms that are related to "machine translation", at different periods in time, in the ACL RD-TEC (Zadeh and Handschuh, 2014). Unfortunately, both studies are based on corpora that cover only very short time spans, namely a mere 15 years in the case of LREC and 41 years in the case of ACL ARC. Furthermore, Mariani et al. (2014) cannot provide a full analysis of terminological dynamics, since the topic of their article is much broader.

Other relevant terminological contributions to the topic are Kristiansen (2011) and Picton (2011). Kristiansen (2011) provides an in-depth analysis of conceptual dynamics in three different domains, whereas Picton (2011) presents a fine-grained typology of diachronic term development patterns. However, none of the two authors outlines a quantitative methodology for the study of the research problem posed.

By annotating terminology in the RSC, we not only hope to provide a methodological baseline for term annotation in diachronic corpora. We also hope to stimulate empirical research into the diachronic development of terms as linguistic units and, in particular, the following research questions:

- What are the factors motivating terminological dynamics?
  - What motivates *term-related* processes such as term formation, terminologisation, the occurrence of variants, or term consolidation?
  - How do *conceptual* dynamics affect terminologies, e.g. paradigm change within a given domain or progress in research?
- What are the parameters by which terminological dynamics can be described?
- How do terms change linguistically over time, especially in comparison to the lexicon of general language?

3. Terminology Annotation

3.1. Preparatory Work

Before beginning with the actual annotation, a training phase was used to draft, test, and finalise the annotation guidelines. The guidelines were written by a person with several years of experience in corpus-based terminology. This person also acted as lead annotator during the whole project. The annotation guidelines explain fundamental concepts of terminology and describe the annotation workflow. Semantic classes that typically cover a relevant part of terms in the analysed domain are described (e.g. materials, methods, machinery and components of machinery, physical phenomena, such as "pressure" or "density", etc.). Moreover, the annotation guidelines provide directions on how to distinguish terms from general language words and establish rules for dealing with several problematic cases, including:

- high-level scientific term candidates such as *radius* or *diameter* should not be annotated,
- complex term candidates such as *Charge of Powder* or *length of the Bore* should only be annotated if they refer to a single complex concept,
- terminological verbs such as *to fire a gun* are included into the annotation,
- proper names should not be annotated,
- generic nouns, such as *method* or *contrivance*, should not be annotated,
- spelling variants and misspelled terms are annotated as long as they are recognizable,
- shortened forms, such as contextual synonyms and variants, should not be annotated.

During the training phase, six full texts covering the whole time span of the RSC were annotated independently by three annotators. After the completion of each text, meetings were held to discuss problematic cases and refine the guidelines. The final annotations were carried out independently by two annotators, namely the lead annotator and a computational linguist working on the same project.

3.2. Data Selection and Annotation Workflow

The RSC covers more than 200 years of English scientific writing. For our purposes, we split this period into five sub-periods. Moreover, it was necessary to obtain a rough division of the data into scientific disciplines. Firstly, there was reason to believe that, in this already non-trivial task, annotation quality would decrease if annotators were forced to annotate texts not only from different centuries, but even from different scientific disciplines. Secondly, inter-disciplinary variation as one motivating factor for differences between terminological structures needed to be excluded from the analysis.

Since it was not feasible to manually annotate the corpus with scientific disciplines, an unsupervised method was adopted for this task. More precisely, we used topic modeling as implemented in MALLET (McCallum, 2002) to annotate texts with preliminary topics for the targeted selection of disciplinary homogeneous texts. A model with 24 topics was found empirically to provide reasonable results (Fankhauser et al., accepted, give a more detailed discussion on the building and use of topic models on this specific data set). Based on this model, we chose texts from the topic "Mechanical Engineering", which is characterised by the words *made, length, weight, end, diameter, iron, instrument, experiments, brass, part, point, line, distance, equal, scale, bar, side, fixed,* and *half.* Furthermore, attention was paid to selecting texts of a relatively short length since the test annotation had shown that text length has a negative impact on annotation consistency. Another constraint was that the samples taken for the five time periods should be comparable in terms of size. Table 1 shows how many texts and how many word tokens were selected for each time period. The number of annotated term occurrences is also given. Moreover, for a full evaluation of the scope of our annotation work, the last row of the table gives the overall number of tokens available for each of the five sub-periods.

Period	1665–1699	1700–1749	1750–1799	1800–1849	1850–1869
# Terms	2,123	2,217	2,473	2,095	1,521
# Texts	20	20	19	13	5
# Tokens	29,973	32,977	32,884	27,174	20,583
# Tokens (RSC)	2,953,048	4,239,163	7,593,971	11,399,195	8,728,191

Table 1: Data selected for term annotation

We used WebAnno (Yimam et al., 2013) for annotation and a series of Python scripts for data manipulation. In order to keep a close eye to annotation quality, meetings were held after the completion of each group of two texts to discuss annotation conflicts. More precisely, we dealt with two types of conflicts:

1. *mismatch of term identification*: annotators mark different, disjoint strings of word tokens;
2. *mismatch of term span definition*: annotators mark different, but overlapping strings of word tokens.

Term annotation is a highly difficult task as evidenced by the controversial discussion in the cited literature. In this situation, annotation decisions can be related to objective "termhood" in two ways: One way is to consider as terms only those tokens that were annotated by a relevant number of annotators (i.e. the majority); another way is to constrain annotation by very detailed guidelines. In our task, deciding whether word tokens are terms or not based on majority annotations was not feasible in our setting with only two annotators. However, even annotation decisions constrained by extremely detailed guidelines can by no means be generalised to represent "objective" termhood as long as the annotations are not performed on a massively quantitative scale. We believe that the solution to this problem is not the optimisation of inter-annotator agreement, but the transparency of annotations. We therefore decided not to normalise annotation conflicts of the first type. Instead, term candidates were marked by confidence attributes as explained in Table 2.

Attribute	Explanation
<term_confidence="L">	Type 1 annotation conflicts
<term_confidence="M">	Type 2 annotation conflicts
<term_confidence="H">	Consensual annotations

Table 2: Term confidence attributes

Figure 1 exemplifies a type 2 annotation conflict (mismatch of term span definition). In the example, one annotator annotated "Ivory Cap", whereas the second annotator annotated only "Ivory". These conflicts are resolved interactively with the help of a Python script. After discussion of the conflict, a new annotation is created by entering the correct span, here "Ivory Cap" (variant A).

Annotator A: C is the under Part of the Ivory Cap .
Annotator B: C is the under Part of the Ivory Cap .

Figure 1: Example of a type 2 annotation conflict

Figure 2 provides an example concordance of our annotations after encoding into the IMS Open Corpus Workbench (Evert and Hardie, 2011). Using mark-up, the example shows type 1 annotation conflicts and consensual annotations. Several term candidates (highlighted in red), including three verb forms, have been annotated by only one annotator and, consequently, are labelled as low-confidence terms. In the same figure, high-confidence terms (highlighted in blue) were annotated by both annotators consensually and, consequently, bear the label "H".

3.3. Annotation Results

Having annotated all periods, we calculate inter-annotator agreement as the F-measure for term spans, following the approach used for the NEGRA corpus (Brants, 2000). Table 6 details the results. The second column gives the obtained agreement per text. The remaining columns indicate how many term occurrences were annotated per confidence level.

10,429 term occurrences were annotated and an average agreement of 0.655 over all annotated texts was achieved. Given the complexity of the task and the fact that our evaluation method calculates a rather conservative measure for agreement, we consider this score acceptable. For example, many of the low-confidence terms, which make up roughly 40% of all annotated term occurrences, are frequently occurring lexical units that are systematically annotated by one of the annotators but not by the other. Every single occurrence of such a token is considered an annotation mismatch. This is further evidenced by the observation that

This Material being gotten in its proper Season , it must be very well dried in the Sun , and more than <term_confidence H>Bark</term_confidence> ; then housed dry , and kept dry for Use ; and when it is to be used , the greater <term_confidence L>Wood</term_confidence> may be <term_confidence L>shaved</term_confidence> small , or cleft fit for the Engine , by and by to be described ; and the smaller to be <term_confidence L>bruised</term_confidence> and cut small by the same Engine : Which done , it must again be dried very well upon a <term_confidence L>Kiln</term_confidence> , and then <term_confidence L>ground</term_confidence> , as <term_confidence H>Tanners</term_confidence> usually do their <term_confidence H>Bark</term_confidence> .

Figure 2: Example concordance visualising mark-up for different confidence levels

agreement is unevenly distributed across texts: Some of the texts contain few controversial tokens, whereas in others, there are many. Consequently, the agreement of individual texts ranges from 0.376 to 0.933.

Text ID	A1	A2	Ratio
102150	5	16	3.2
101154	38	111	2.9
101232	46	130	2.8
105866	100	50	0.5
107772	173	85	0.5
107938	44	19	0.4

Table 3: Number of terms annotated by each annotator in texts with low agreement

	1665–1699	1700–1749	1750–1799	1800–1849	1850–1869
Agreement	0.66	0.66	0.70	0.59	0.62
High-confidence terms	51%	54%	60%	47%	50%
Low-confidence terms	46%	41%	34%	45%	45%

Table 4: Annotation quality for different periods

3.4. Reasons for Disagreement

Although the degree of agreement achieved in our first run of annotations is acceptable, it seems reasonable to think of ways how annotations can be made more reliable. In terminology annotation projects, linguistic annotation guidelines are normally used to constrain annotation decisions and to optimise agreement. Detailed instructions are given for semantic categories and candidate token sequences. One major difficulty in the development of such guidelines is that it is virtually impossible to know in advance which candidate strings will be the most difficult to annotate.

As suggested before, we do not believe that the elaboration of even more detailed linguistic guidelines can constrain term annotation decisions to a sufficient degree: In fact, a lexical unit can be a term in one context or text but not in another. The guidelines underlying our annotations are already quite detailed, but they resulted in modest rather than high agreement. We hypothesize that the experience of individual annotators as well as other personal and textual factors influence annotation performance. For example, less experienced annotators seem to annotate more terms.

In what follows, we therefore present an analysis of our annotations, based on which we formulate proposals how annotation quality could be improved in future settings. Table 3 provides an overview over the number of term occurrences that were annotated by each of the annotators in a selection of texts in the current project stage. A1 is the more experienced annotator. The last column gives the ratio of the numbers for A1 and A2. The selected texts are extreme cases, namely the three texts with the highest ratios and the three texts with the lowest ratios. Overall, in 50 out of 77 texts, annotator A2 annotated more term occurrences than annotator A1, the average ratio being 1.2. Hence, annotator experience indeed seems to influence annotator behaviour.

Another aspect that seems to affect annotation quality is text length: Longer texts are lexically more varied, and it is harder for annotators to annotate consistently in very long texts than in shorter ones. Unfortunately, in the RSC, texts from earlier periods are relatively short, whereas they tend to be quite long in later periods, and this affected our choice of annotation material. Moreover, as can be seen from Table 4, this also affected our annotation results. The table provides agreement scores averaged over the respective time periods. As another quality indicator, the second and third row of the table provide the percentage of high-confidence and low-confidence terms.

Table 4 supports our hypothesis that annotation becomes more difficult in the 1800–1849 period, for which a smaller set of rather long texts had been selected. For the last period, quality seems to recover a bit, but it remains considerably worse than in the earlier periods. This finding is actually quite surprising because language use in the more recent texts is closer to modern standards and thus, in theory, these texts should be easier to annotate. In summary, our results suggest that future annotation projects should concentrate on shorter texts or even text snippets instead of full texts.

As a step towards enhancing the reliability of future work on the same data set, we also tried to spot those term candidates that were particularly controversial in the current annotation stage. As pointed out before, it is hardly possible to identify such term candidates before the actual annotation work begins. However, it is feasible to elaborate methods for the identification of spurious candidates in already existing annotations. This information can then be used to either clean the annotated data or as a basis for the refinement of the annotation guidelines for future annotation rounds. For the identification of spurious term candidates in our data, we applied a sequence of simple steps:

1. Select frequently annotated terms, i.e. terms annotated at least 20 times.

2. For each frequently annotated term, count how many of the annotated occurrences have low confidence.

3. Finally, select those terms that have low confidence in at least 80% of the annotations.

This method revealed a list of quite unreliable term candidates, such as *deviation* (1 high-confidence annotation, 21 low-confidence annotations) or *rod* (5 high-confidence, 1 medium-confidence, and 67 low-confidence annotations).

All in all, our analysis of agreement and disagreement shows that term annotation performance is influenced both by personal annotator characteristics as well as features of the annotated texts themselves. Factors such as text length and annotator experience need to be controlled and methods for detecting spurious annotations have to be devised and applied. In terms of annotation guidelines, a more fine-grained distinction between different types of terms (e.g. topic keywords, scientific standard vocabulary, foreign language words, unknown or “strange” words, . . .) might be helpful, at least partly, in alleviating the difficulty of the annotation task.

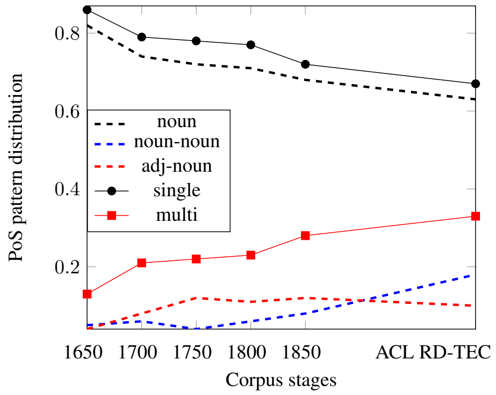

Figure 3: Development of PoS patterns over time

	1650	1700	1750	1800	1850	ACL RD-TEC
Term length						
single-word	6.19	6.46	6.53	6.65	7.10	7.6
overall	7.31	8.16	8.41	8.57	9.68	10.9

Table 5: Term length in characters over time

4. Properties of Annotated Terms

We exploit our annotation results for studying term surface features, using terminology annotated in the ACL Anthology Reference Corpus (ACL RD-TEC) (Zadeh and Handschuh, 2014) for comparison with modern data. The surface features are:

- term length in characters,
- term length in words,
- term PoS patterns.

It should be noted that this comparison has its limitations. In our annotation of the RSC, not each occurrence of a given candidate was annotated as a term, but only when considered appropriate by at least one annotator. Our version of the ACL RD-TEC, however, was annotated automatically using the reference list provided by Zadeh and Handschuh (2014) as a dictionary. Consequently, each occurrence of a candidate, in this corpus, *was* annotated as a term if it was in the reference list. Moreover, PoS tagging on the RSC still has its flaws, so the values for certain PoS tags might slightly change in later stages of corpus annotation and analysis. The corpora are also quite unequal in size: the ACL RD-TEC contains 2,822,986 terms, whereas the RSC data is considerably smaller (see Table 1). The analysis, therefore, really is a pilot study aiming at identifying relevant questions for future research rather than providing consolidated information about the diachronic development of terms in English scientific writing.

Figure 3 shows the proportion of occurrence of the 3 most important PoS patterns among both terms annotated in the RSC and term annotations in the ACL RD-TEC. The share of single- and multi-word terms in both data sets is also given. Table 5 summarises the development of word length in characters for both single-word terms and all terms.

What can be seen from the analyses is that term length in words increases more than just slightly in two periods, namely between 1650 and 1700 and between 1800 and 1850, however, without reaching the distribution of the modern data (ACL RD-TEC). We also see that modification of a base term by an adjective (pattern *adj-noun* in Figure 3) is a linguistic resource that is adopted early and seems to remain stable until modern days. Noun-noun compounds (pattern *noun-noun* in Figure 3) appear only in the 19th century, but over time they become far more typical for multi-word term constructions than other means of word formation. The development of term length in characters matches this development, especially the more pronounced growth between 1650 and 1700 as well as 1800 and 1850.

It is difficult to come up with interpretations for the observed development, since, obviously, the findings of our analyses have to be related to parallel processes in general language, which is not an easy undertaking. Nevertheless, we would like to sketch at least a few ideas along the lines put forward in the abstract of this article, namely developments on both the semantic and the morpho-syntactic level.

One line of interpretation is the analysis of morpho-syntactic complexity. Looking at our results from this angle, we observe a quite steady increase in complexity that, with little jumps in the two periods mentioned, at first extends noun phrases analytically (by means of modifying adjectives), and then gives way to denser and less explicit forms (noun-noun compounds). According to Biber and Gray (2011), the latter happens only in the second half of the 19th century, whereas our data indicates a slow, but steady rise of this construction already in the 18th century. The reasons for this observation are by no means clear. While PoS tagging errors might have contributed to the observed effect, another explanation is the influence of other languages in which learned prose at that time was normally written (e.g. Latin or French). Moreover, the development of adjectival modification, too, seems to suggest a pioneering role of terminology with respect to the increase in noun phrase complexity in scientific prose. Again, Biber and Gray (2011) describe an increase in adjectival modification only in the 18th century, whereas, in our data, this process already starts earlier. The results of Biber and Gray (2011) are easily confirmed with the help of large-scale resources such as the Google Books Ngram Viewer[1], however, our study is, to our knowledge, among the first to concentrate on terminology only, and therefore the hypothesis of a pioneering role of terminology in the development of the noun phrase should not be rejected too quickly.

Another aspect that we would like to mention is related to the distance between the lexicon of general language and terminologies. The length of terms in characters increases over time and it seems reasonable to relate this finding to word frequency and, thus, word familiarity. This relation goes back to seminal work by Zipf (1932) (see also (Prün, 2005)) according to whom frequently used words are shorter than less frequently used lexical units. Moreover, this finding actually matches the annotators' intuition who felt that term candidates were frequently taken from the lexicon of general language in the earlier annotation periods and thus quite hard to distinguish from general-language words. Obviously, this hypothesis needs to be further researched, potentially using more advanced techniques such as language modeling, entropy analysis, etc.

5. Discussion and Future Work

In this paper, we have described our methodology for annotating terminology in a diachronic corpus of English scientific texts. 77 texts containing more than 140,000 words tokens were manually annotated, yielding more than 10,000 occurrences of terms. A detailed analysis of our annotation results was provided, including an analysis of disagreement followed by proposals for the improvement of annotation quality. It was argued that not only linguistic, term-related aspects need to be taken into account for obtaining reliable annotations, but also annotator- and workflow-related variables.

We also provided—very preliminary—evidence for our hypothesis that terms are subject to diachronic processes on both the semantic and the morpho-syntactic level. Our data suggests that, in the earlier periods of the corpus, terms are not easily distinguishable from non-terms, that is, they are closely related and actually often directly taken from general language vocabulary. In the course of the roughly 200 years analysed here, terms undergo processes that increase the distance between specialised and everyday language. Moreover, we found evidence that terminology actually plays a pioneering role in noun-phrase related processes leading to the development of linguistic means for efficient and concise communication. Future work will need to verify these preliminary results.

6. Acknowledgements

The work described in this paper was funded by the DFG through the Cluster of Excellence "Multimodal Computing and Interaction". We are also grateful to Elke Teich for supporting the conference presentation of this paper. Moreover, we thank Jörg Knappen for sharing his topic modeling results. Last but not least, we thank our anonymous reviewers for their helpful comments.

7. Bibliographical References

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W., Cohen, K., Verspoor, K., Blake, J., and Hunter, L. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

Banks, D. (2008). The Significance of Thematic Structure in the Scientific Journal Article, 1700-1980. In *Systemic Functional Linguistics in Use*, pages 1–29. University of Southern Denmark.

Banks, D. (2009). Creating a specialized discourse: the case of the Philosophical Transactions. *ASp – la revue du GERAS*, (56):29–44.

Biber, D. and Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics*, (15):223–250.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of LREC'08*, Marrakech, Morocco. ELRA.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3):993–1022.

Brants, T. (2000). Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of LREC'00*, Athens, Greece. ELRA.

Drouin, P. (2004). Detection of Domain Specific Terminology Using Corpora Comparison. In *Proceedings of LREC'04*, Lisbon, Portugal. ELRA.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK. University of Birmingham.

Fankhauser, P., Knappen, J., and Teich, E. (accepted). Topical Diversification over Time in the Royal Society Corpus. In *Proceedings of DH 2016*, Krakow, Poland, July.

Gupta, S. and Manning, C. D. (2011). Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of IJCNLP'11*, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. In *Proceedings of EMNLP'08*, Honolulu, Hawaii. ACL.

Khamis, A., Degaetano-Ortlieb, S., Kermes, H., Knappen, J., Ordan, N., and Teich, E. (2015). Introducing the Royal Society Corpus: A resource for the diachronic study of scientific English. In *Corpus Linguistics 2015 Conference*, Lancaster, UK. University of Lancaster.

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

Kristiansen, M. (2011). Domain dynamics in scholarly areas: How external pressure may cause concept and term changes. *Terminology*, 17(1):30–48.

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014). Rediscovering 15 Years of Discoveries in Lan-

[1] https://books.google.com/ngrams.

guage Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of LREC'14*, Reykjavik, Iceland. ELRA.

McCallum, A. K. (2002). Mallet: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Moessner, L. (2009). The influence of the Royal Society on 17th-century scientific writing. *International Computer Archive of Modern and Medieval English*, (33):65–88.

Picton, A. (2011). Picturing short-period diachronic phenomena in specialised corpora: A textual terminology description of the dynamics of knowledge in space technologies. *Terminology*, 17(1):134–156.

Prün, C. (2005). Das Werk von G. K. Zipf. In *Quantitative Linguistik/Quantitative Linguistics*, pages 142–152. de Gruyter.

Schumann, A.-K. and QasemiZadeh, B. (2015). Tracing Research Paradigm Change Using Terminological Methods: A Pilot Study on "Machine Translation" in the ACL Anthology Reference Corpus. In *Proceedings of TIA'15*, Granada, Spain. CEUR Workshop Proceedings.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of ACL'13*, Sofia, Bulgaria. ACL.

Zadeh, B. Q. and Handschuh, S. (2014). The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Proceedings of the Computerm'14 Workshop*, Dublin, Ireland. ACL.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.

Text ID	Agreement	Confidence=H	Confidence=M	Confidence=L
100986	0.875	28	0	8
101154	0.443	33	3	78
101211	0.872	112	1	31
101232	0.42	37	3	96
101314	0.62	66	2	77
101363	0.747	56	5	30
101435	0.442	17	1	41
101739	0.619	61	3	71
101777	0.757	56	3	30
101812	0.725	74	5	48
102071	0.598	134	15	150
102072	0.584	33	11	25
102119	0.762	56	2	32
102150	0.476	5	0	11
102162	0.635	27	6	19
102281	0.604	64	2	80
102326	0.796	45	2	19
102360	0.672	46	4	37
102500	0.766	54	1	31
102567	0.71	76	2	58
103423	0.64	48	14	24
103433	0.747	31	0	21
103510	0.704	82	4	62
103527	0.836	84	1	31
103530	0.865	16	1	3
103792	0.71	76	3	57
103842	0.615	71	9	73
103891	0.796	127	6	54
103948	0.525	31	7	36
104074	0.443	27	5	58
104104	0.636	62	14	44
104191	0.376	19	3	57
104242	0.702	40	0	34
104257	0.618	51	0	63
104322	0.68	103	12	73
104389	0.782	70	4	33
104441	0.593	51	9	52
104686	0.631	76	17	54
104800	0.795	89	6	34
104802	0.569	39	9	37
105019	0.685	37	2	30
105085	0.893	129	3	25
105095	0.645	40	10	25
105259	0.598	35	4	41
105322	0.933	194	9	10
105346	0.744	80	5	45
105519	0.783	130	17	39
105866	0.413	31	4	79
105879	0.81	64	7	15
105944	0.615	88	21	68
105962	0.773	85	4	42
106045	0.745	38	2	22
106066	0.599	41	2	50
106178	0.65	66	5	61
106424	0.775	93	9	33
106435	0.488	71	37	77
106468	0.658	73	5	68
106676	0.707	100	18	48
106765	0.751	83	2	51
107265	0.673	75	16	41
107397	0.582	41	12	39
107398	0.647	43	6	35
107403	0.636	154	35	107
107461	0.489	43	6	78
107621	0.593	97	7	119
107684	0.669	112	19	80
107720	0.525	69	7	111
107736	0.695	130	14	80
107772	0.426	55	23	91
107930	0.614	58	9	57
107938	0.476	15	1	31
108196	0.699	102	11	66
108916	0.65	115	13	97
108919	0.561	32	4	41
108922	0.689	204	23	137
108944	0.572	117	12	127
108972	0.625	287	28	284
Overall	0.655	5,500	607	4,322

Table 6: Agreement and number of terms per text