THE BULLETIN OF OKAYAMA UNIVERSITY OF SCIENCE No. 35 B pp. 47-54 (1999)

Structure and Linguistic Features of Research Article Titles in Computer Science

Laurence ANTHONY

Department of Information and Computer Engineering,
Faculty of Engineering,
Okayama University of Science,
Ridai-cho 1-1, Okayama 700-0005, Japan

(Received November 4, 1999)

This paper describes the structure and linguistic features of research article titles in computer science, focussing specifically on preposition usage, length, punctuation usage, and word frequency. The research is based on a corpus of 408 articles, which represents all the research articles published in the six journals of the IEEE Computer Society in 1998. Results show that there is a wide variation in title length across the six journals. This suggests that title length is not discipline dependent, and rather, dependent on the nature of the study or problem investigated. Punctuation in titles is shown to be restricted to the exclusive use of colon constructions, although their frequency of usage varies depending on the journal. The highest frequency words in the titles are prepositions and articles, with 'for' appearing as the most or second most frequent word in all journals. Other high frequency words vary depending on the journal, and reflect the nature of its content. Finally, titles are written using a very compact structure, removing unnecessary phrases such as 'Investigation on' or 'Studies on', and words such as 'a', 'an' and 'the' in the opening position.

Introduction

Many writers have commented on the importance of the title in the decision to read a research article (RA) (e.g. Bazerman, 1985; Day, 1994, Taniguchi et. al, 1995). Day (1994:15), for example, points out that the "title will be read by thousands of people", although "few people, if any, will read the entire paper". The difficulty in writing an effective title has also been commented on. For example, Swales (1990:222) points out that "composing the few words of a title can take up an inordinate amount of time, discussion and mental effort". Surprisingly, however, the amount of research that has focussed on titles is relatively scarce, especially compared with the work done on other parts of the RA, such as the introduction and discussion sections. Also, when titles have been looked at, the discussions have tended to be short and more intuitively based (e.g. Swales et. al, 1994; Kinoshita, 1996, Nakajima et. al, 1996). One noticeable exception to this is the recent work of Fortanet et al. (1997, 1998) and Posteguillo (1998), who have published several papers on various aspects of title writing based on an analysis of articles published in linguistics, business and economics, chemistry, and computer science journals.

Continuing from the work of Fortanet et al. (1997), this paper describes the structure and linguistic features of RA titles in computer science, focussing specifically on preposition usage, length, punctuation usage, and word frequency. The research is based on a corpus of 408 articles, which represents all the RAs published in the six journals of the IEEE Computer Society in 1998. Due to the shear size of the corpus, it was necessary to automate part of the analysis, and the tools and techniques used will be described. Many of the results here can be compared directly with those obtained by Fortanet et. al (1997), although some differences emerge. Possible reasons for these are offered in terms of corpus design and sub-discipline characteristics.

Corpus Selection and Preparation for Analysis

In previous research projects of this kind, the corpus has been selected using various different criteria. For example, some researchers have chosen to look at the first articles published in each edition of a journal over several months, while others have looked at all the articles published in a single journal edition. Others still, have made a totally random selection or selected articles which show specific features (e.g. Tarone et al., 1981; Swales, 1981; Cooper, 1985). For example, in a previous study, we chose a corpus of 12 research articles in a single journal that had been awarded 'best paper' awards (Anthony, 1999)

Unfortunately, all these selection procedures have drawbacks. A selection based on articles with certain features clearly affects the generalizability of results. On the other hand, a totally random selection or a selection based on the first articles of a journal edition has the danger of including non-representative texts of a journal. The only way to avoid this is to use a very large corpus, a technique which has been used very rarely to date.

In this study, in order to generate results which accurately represent the field of computer science with high generalizability, we decided to prepare a large corpus of RAs from the field's leading journals. Therefore, we chose to create a corpus comprising all the research articles published in the six journals of the IEEE Computer Society, the largest computer society in the world. The journal titles and their respective number of articles are listed in Figure 1. Note that abbreviations of the journal titles appear in parentheses, and will be used throughout the remainder of the paper.

Figure 1 Corpus Journal Titles and Number of Articles

Journal Name	**Number of articles**
1) *Transactions on Computers (C)*	89
2) *Transactions on Knowledge and Database Engineering (KDE)*	51
3) *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*	72
4) *Transactions on Parallel and Distributed Systems (PDS)*	97
5) *Transactions on Software Engineering (SE)*	73
6) *Transactions on Visualization and Computer Graphics (VCG)*	26
Total	408

Due to the size of the corpus, it was necessary to prepare the corpus in a text file format for automatic analysis by computer. Fortunately, the IEEE Computer Society publishes electronic versions of their journals so it was possible to download the articles from the IEEE Computer Society internet site, avoiding the need to scan texts into the computer, followed by an often frustrating process of using OCR software to convert the images to text format[1]. One problem with the electronic versions of the journals was that they were in PDS file format. However, by using an extension to Adobe Acrobat Reader it was possible to open the files and save them in the desired format[2].

After the entire texts of the RAs were saved as text files, it was necessary to extract the titles of the RAs and save these as separate files. This part of the corpus preparation was conducted by hand, although a multitude of keyboard shortcuts reduced the time considerably. Finally, each file was named enabling the source text to be easily reference, and grouped in a separate folder for each journal.

Methodology

The features of the RA titles to be investigated were length, punctuation usage, word frequency, and structure. First, title length and word frequencies were calculated using the concordance program Wordsmith 3.00, developed by Mike Scott of Liverpool University[3]. These results were exported to a Microsoft Excel spreadsheet to calculate averages, and plot the various frequency distributions. Titles which included punctuation marks were located in the corpus using a simple search program[4]. The relationships between the sections of the titles separated by punctuation marks were then analyzed by human observation. To improve the reliability of

this type of analysis, three specialists in the field of computer science were consulted. This was particularly important as it was sometimes difficult to categorize title sections that assumed specialist knowledge of the subject matter. Finally, the structure of titles was investigated by human observation, again through consultation with the three specialist informants.

Results and Discussion

Length of RA Titles

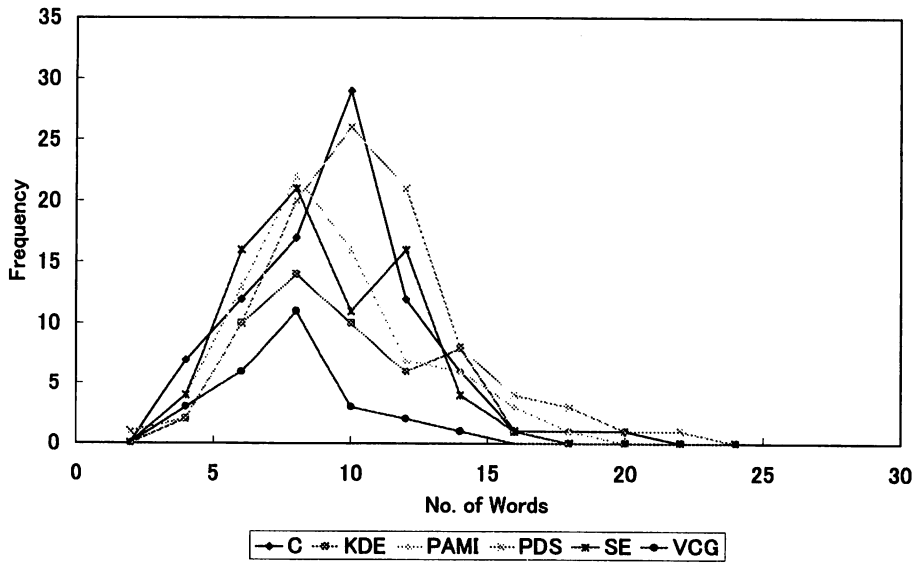
The maximum, minimum and average length of the corpus RA titles is shown in Table 1. As can be seen , there is a wide variation in title length across all six journals, with an average title length variation of 7.4 to 10.0 words. A clearer picture of title length variation can be gained from a frequency distribution plot of title lengths. This is shown in Fig. 2. Again, the wide variation in length across journals can be seen, although we see two frequency peaks in the cases of KDE and SE journals, suggesting a preference for two different types of title structure.

The results differ slightly from those of Fortanet et. al (1997), who reported a title length variation in computer science RAs of 4 to 8 words, with an average of 7.32 words. Two possible reasons account for this: 1) Fortanet et. al used a corpus of 50 RAs from five different journals. Assuming only ten articles were selected from each journal, it is unlikely they would provide a representative sample of the journal as a whole. 2) Fortanet et. al do not state which journals were investigated, suggesting that perhaps the writers in the journals here have more freedom in the choice of titles. Note that here, no directives on the writing or length of titles are issued by the journals[5]. One further point is that without a frequency distribution plot such as that in Fig. 2, it is difficult to interpret results such as the maximum and minimum, and especially averages. For example, an average length does not reveal the tendency to use two different title lengths, as in the case of KDE and SE articles here.

Table 1 Length of RA Titles

No. of Words	C	KDE	PAMI	PDS	SE	VCG	Overall Ave.
Min.	3	4	3	2	4	3	3.17
Max.	20	15	17	21	16	14	17.17
Ave.	9.0	9.0	8.7	10.0	8.4	7.4	8.75

Figure 2 Frequency Distribution Plot of RA Title Lengths



Punctuation Usage in RA Titles

In contrast to other disciplines such as applied linguistics, which use a variety of punctuation marks in their titles, in the corpus RA titles with punctuation the authors used the colon exclusively. There was, however, a noticeable difference in usage across the journals, with approximately 8% of C and VCG titles using punctuation compared with almost 20% of KDE titles. See Table 2. These results are consistent with those of Fortanet et. al (1997) who also report on the exclusive use of colon punctuation in computer science titles, accounting for 12% of the corpus articles.

Table 2 Punctuation Usage and Relationships between Sections of RA Titles

% Use	C	KDE	PAMI	PDS	SE	VCG	Total
Colon Constructions	8.0	19.6	15.3	11.3	17.8	7.7	13.3
Name: Description	3.4	9.8	2.8	7.2	6.8	7.7	5.9
Description: Name	1.1	0.0	0.0	0.0	0.0	0.0	0.2
Topic: Description	0.0	2.0	1.4	0.0	0.0	0.0	0.5
Topic: Scope	3.4	7.8	8.3	1.0	5.5	0.0	4.4
Topic: Method	0.0	0.0	2.8	3.1	5.5	0.0	2.2

In order to investigate the relationship between the sections of the RA separated by the colon punctuation, Fortanet et. al (1997a) use the categories proposed by Swales et al. (1994), i.e., problem-solution, general-specific, topic-method, and major-minor. From an initial analysis of the titles here, it was felt that these categories did not capture the essence of the title sections, and so a new set was developed. These are listed in Table 3 below, and the results of the analysis are shown in Table 2. It can be seen from Table 3 that there is considerable overlap with the categories suggested by Swales et al. (1994), and although the problem-solution relationship is missing (as it was not observed in the corpus titles) it is anticipated that this would appear in other RA titles. One major difference, however, is that there is no assumption here that one category would necessarily appear with another. For example, although we would expect the "Scope" category to follow "Topic", evidence shows that this is not always not case. In this respect, the categories resemble those offered by Hamp-Lyons (1987), who describes titles as being composed of topic, focus, comment and viewpoint, although not all elements are expected to appear in each title.

Table 3 Categories Used in Labeling RA Title Sections Separated by Colon Punctuation

Name of Approach/Algorithm/Application etc.
Description of Approach/Algorithm/Application etc.
Topic of Research Article
Scope of Research Article
Method of Research

Table 2 shows that the 'Name: Description' relationship is the most common way to structure punctuated titles, followed by the 'Topic: Scope' relationship. For example,

'Name: Description'
"MPS: Miss- Path Scheduling for Multiple- Issue Processors" (C)
"ADOME: An Advanced Object Modelling Environment" (KDE)
"Veinerization: A New Shape Description for Flexible Skeletonization" (PAMI)
"Macro- Star Networks: Efficient Low- Degree Alternatives to Star Graphs" (PDS)
"KLAIM: A Kernel Language for Agents Interaction and Mobility" (SE)
"The Information Mural: A Technique for Displaying and Navigating Large Information Spaces" (VCG)

'Topic: Scope'
"Cache Prefetching: Timing Evaluation of Hardware Implementations" (C)
"Collaborative Multimedia Systems: Synthesis of Media Objects" (KDE)
"Fingerprint Image Enhancement: Algorithm and Performance Evaluation" (PAMI)
"The CLAM Approach to Multithreaded Communication on Shared- Memory Multiprocessors: Design and Experiments" (PDS)
"Coordinating Multiagent Applications on the WWW: A Reference Architecture" (SE)

However, there is a wide variation in their usage across the different journals. For example, although they are used almost evenly in C, KDE and SE journals, PDS titles show a strong preference for the former whereas PAMI titles show a preference for the latter. VCG titles, on the other hand, only exhibit the 'Name: Description' relationship, although the sample size here is relatively small. There is also a variation in usage of the 'Topic: Method' relationship. SE titles show a strong preference for its usage, for example,

"Communication and Organization: An Empirical Study of Discussion in Inspection Meetings"
"Modelling and Evaluating Design Alternatives for an On- Line Instrumentation System: A Case Study"

However, there are no examples in C, KDE and VCG titles.

It is suggested that the differences in punctuation usage reflect the nature of research in the different journals. For example, much of the research published in the PDS journal describes a new system or approach, hence we find many cases where the name of the developed approach or algorithm is followed by its description.. For example,

"Spanning Multichannel Linked Hypercube: A Gradually Scalable Optical Interconnection Network for Massively Parallel Computing"
"The Offset Cube: A Three- Dimensional Multicomputer Network Topology Using Through- Wafer Optics"

Research in the PAMI journal, on the other hand, is dealing with more general problems and so we often see a description of a general topic followed by the scope of the present article. For example,

"Junctions: Detection, Classification, and Reconstruction"
"Partial Classification: The Benefit of Deferred Decision"

Word Frequency and Structure of RA Titles

It was anticipated that the greatest variation across the journal titles would be in their word frequency. This is because each journal could be said to represent a different sub-discipline of computer science, with its own set of concepts and technical terms. Table 4 shows a list of the 20 most frequent words which appear in each journal. From the table, it can be seen that there is considerable variation across the journals reflecting the nature of each sub-discipline. For example, PAMI titles show a high frequency of words related to machine intelligence such as 'recognition', 'detection', 'learning', and 'image', whereas VCG titles show a high frequency of words related to computer graphics, such as 'geometric', 'line', 'meshes' and 'ray'.

The highest frequency words in all the journals, however, were prepositions and articles. In particular, the preposition 'for' appears as the highest frequency word in four journals and the second highest in the remaining two. This result reflects the nature of engineering which is to develop something which is useful; something that can be applied for some specific purpose. For example,

Table 4 Highest Frequency Words Appearing in RA Titles

	C	KDE	PAMI	PDS	SE	VCG	Overall
1	FOR	FOR	AND	A	FOR	FOR	FOR
2	OF	IN	FOR	FOR	OF	OF	OF
3	AND	AND	OF	IN	A	A	A
4	A	A	A	AND	AND	AND	AND
5	IN	DATABASES	USING	ON	IN	EFFICIENT	IN
6	SYSTEMS	OF	RECOGNITION	OF	SOFTWARE	USING	THE
7	THE	TEMPORAL	THE	THE	THE	VOLUME	ON
8	CACHE	DATA	TO	PARALLEL	USING	COLLISION	USING
9	USING	OBJECT	IN	DISTRIBUTED	REQUIREMENTS	DATA	TO
10	NETWORKS	PROCESSING	ANALYSIS	TIME	TO	DETECTION	SYSTEMS
11	OPTIMAL	SYSTEMS	BASED	ALGORITHM	ANALYSIS	FAST	WITH
12	WITH	THE	DETECTION	SYSTEMS	DESIGN	FREE	ANALYSIS
13	FAULTS	EFFICIENT	LEARNING	TO	AN	GEOMETRIC	BASED
14	ON	KNOWLEDGE	ON	EFFICIENT	ENGINEERING	HIERARCHIES	PARALLEL
15	PARALLEL	ORIENTED	WITH	NETWORKS	BASED	INFORMATION	AN
16	ANALYSIS	TO	APPROACH	MESHES	USE	LINE	DISTRIBUTED
17	BASED	WITH	FROM	SCHEDULING	DEVELOPMENT	MESHES	ALGORITHM
18	FAULT	AN	IMAGE	USING	SCENARIOS	MODELS	EFFICIENT
19	LEVEL	ANALYSIS	ALGORITHM	WITH	SYSTEM	MULTIRESOLUTION	DATA
20	PERFORMANCE	APPROACH	AN	ALGORITHMS	SYSTEMS	RAY	TIME

"Abstraction techniques for validation coverage analysis and test generation" (C)
"Designing access methods for bitemporal databases" (KDE)
"Example-based learning for view-based human face detection" (PAMI)
"Abstractions for portable, scalable parallel programming" (PDS)
"Compositional programming abstractions for mobile computing" (SE)
"A high accuracy volume renderer for unstructured data" (VCG)

Another use of prepositions is to narrow the focus of the article. We see this most profoundly in some of the very long titles which appear in the corpus. For example,

"Performance evaluation and cost analysis of cache protocol extensions for shared-memory multiprocessors" (C)
"A multiagent update process in a database with temporal data dependencies and schema versioning" (KDE)
"A volumetric/iconic frequency domain representation for objects with application for pose invariant face" (PAMI)
"A priority-driven flow control mechanism for real-time traffic in multiprocessor networks" (PDS)
"Existence dependency: The key to semantic integrity between structural and behavioral aspects of object types" (SE)
"Fast horizon computation at all points of a terrain with visability and shading applications" (VCG)

This suggests that length variation is dependent not on the journal or sub-discipline, but on the type of study or problem being investigated. This is highlighted by the fact that both the shortest and longest titles in the corpus come from the same journal, PDS.

"Diskless checkpointing"
"Collection-aware optimum sequencing of operations and closed-form solutions for the distribution of a divisible load on arbitrary trees"

Day (1994:15) defines an RA title as "the fewest possible words that adequately describe the contents of the paper". He then goes on to criticize authors who use 'waste' words, in particular phrases such as 'Investigation on' and 'Observations on', and an opening 'A', 'An' or 'The' in the title. His observations seem to have been heeded by the authors of the journal articles here. Although 'a', 'an' and 'the' appear with a high frequency in

the titles, they a rarely found in the opening position, and no article in the corpus opens with 'Investigation on' or other similar phrase. Rather, the terms in the titles appear to have been carefully selected to represent as comprehensively as possible the content of the articles, a technique strongly advocated by Day. One reason for this is that it enables the articles to be effectively retrieved from abstracting and indexing publications, and internet search engine databases.

Conclusion

Fortanet et al. (1997) conclude that titles of RAs vary in length depending on the discipline. From the results here, we can add a post-modifier to this, i.e., titles of RAs in a single discipline vary in length depending more on the problem or nature of the research than on the sub-discipline being examined. The use of punctuation in computer science RA titles is shown to be in general rare, and when it is used there is a very strong tendency to use colon constructions. Here, however, the decision to use punctuation and the type of relationship between the different sections of the title appear to be discipline dependent.

As expected, the highest frequency words in the RA titles were prepositions and articles. The role of prepositions in the titles was particularly important as they were used to narrow the scope of the research topics, and/or specify the precise area in which the developed tool or approach could be applied. Other high frequency words were mainly nouns which reflect the nature of the different journal content.

Finally, the titles showed a tendency to avoid using words or phrases with low content value, enabling the articles to be easily retrieved from abstracting and indexing publications, and search engine databases.

Notes

[1] The current address of the IEEE Computer Society internet site is http://www.computer.org/.

[2] Adobe Acrobat Reader can be downloaded from the internet at the address http://www.adobe.com/. The plug-in used to enable PDS files to be saved as text files was Adobe Acrobat Access 1.0B2, which can be downloaded from the same address.

[3] Wordsmith 3.00 can be downloaded from the internet at the address http://www1.oup.co.uk/elt/catalogu/

[4] The search program used was Windows Grep 2.1.2, which can perform searches using regular expressions. The program can be downloaded from the internet at the address http://www2.pncl.co.uk/~huw/grep/

[5] Notes for authors submitting articles to journals can be usually found on the last page of the journal's first edition of the year. Additional information is supplied in the IEEE guidebook "Information for IEEE Transactions, journals and letters authors', which is available from IEEE Operations Center, Transactions/Journals Department, PO Box 1331, Piscataway, NJ 08855-1331.

References

Anthony, L. (1999). Writing research article introductions in software engineering: How accurate is a standard model? *IEEE Trans. Prof. Commun.* Vol. 42:1, pp. 36-46.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. W*ritten Communication.* Vol. 2:1, pp. 3-23.

Cooper, C. (1985). *Aspects of article introductions in IEEE publications.* Unpublished M.Sc. dissertation. University of Aston, UK.

Day, R. A. (1994). *How to Write and Publish a Scientific Paper.* Cambridge: CUP.

Fortanet, I., Coll, J. F., Palmer, J. C. and S. Posteguillo (1997). The writing of titles in academic research articles. In R. M. Chamorro and A. R. Navarrete (eds.) L*enguas Aplicadas a las Ciencias y la Tecnologia Approximaciones.* Caceres: Universidad de Extremadura, Servicio de Publicaciones.

Fortanet, I., Posteguillo, S., Coll, J. F., and J. C. Palmer (1998). Linguistic analysis of research article titles: Disciplinary variations. In I. Vazquez and I. Camilleu (eds.) *Perspectivas Praguietices en Linguistics Aplicada.* Zaragozs: Anubar.

Hamp-Lyons, L. (1987). *Study Writing: A course in written English for academic and professional purposes.* Cambridge: CUP.

Kinoshita, K. (1996). *Technical writing in Science* (in Japanese). Tokyo: Chuo Koronsha Press.

Nakajima, T and S. Tsukamoto (1996). *Writing Intelligent Scientific and Technical Texts.* Tokyo: Corona Press.

Posteguillo, S. (1998). Writing titles for computer science research articles: Sub-disciplinary variations in academic English. In J. T. Lundquist, H. Picht and J. Quistgaard (eds.) *Proceedings of the 11th European Symposium on Language for Special Purposes: LSP, Identity and Interface, Research, Knowledge and Society.* Copenhagen: Copenhagen Business School.

Swales, J. (1981) *Aspects of article introductions.* Birmingham, UK: The University of Aston, Language Studies Unit.

Swales, J. M. (1990). *Genre Analysis: English in academic and research settings.* Cambridge: CUP.

Swales, J. M. and C. B. Feak (1994). *Academic Writing for Graduate Students.* Ann Arbor: Univ. of Michigan Press.

Taniguchi, S., Tanaka, T., Iida, T. and J. D. Cox (1995). *Writing scientific and technical papers in English* (in Japanese). Tokyo: Chuo Press.

Tarone, E., Dwyer, S. Gillette, S. and V. Icke (1981). On the use of the passive in two astrophysics journal papers. *The ESP Journal.* Vol. 1, pp. 123-140.