THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES
Vol. 3, Nº 2, 2015, pp. 295−304

UDC 811.111'276.6:621.38

CREATING AN ENGINEERING ACADEMIC FORMULAS LIST

Jessica Fox, Magda Tigchelaar

Michigan State University, USA
Phone: 231.245.6905 Email: foxjess3@msu.edu
Phone: 616.214.0823 Email: tigchel1@msu.edu

Abstract. *This study presents a partial replication of the Academic Formulas List (AFL) project (Ellis, Simpson-Vlach & Maynard 2008). The objective of the present study was to identify a corpus derived, pedagogically useful list of formulaic sequences for technical writing in engineering, called the Engineering Academic Formulas List (EAFL) by triangulating corpus metrics with engineering instructors' insights.*
This list of formulas was created using the following criteria: (1) highly frequent, recurrent formulas that were extracted from a 1 million word corpus of published engineering research articles, which (2) occurred significantly more often in the engineering corpus than a corpus of 1.5 million words of general academic discourse, and (3) appeared in a wide range of engineering subfields and publications. Approximately 765 formulas fit these criteria (e.g., a function of the). Next, to determine which of these formulas were most pedagogically useful, 12 graduate level engineering teaching assistants rated whether the formulas extracted from the expert texts were worth teaching to newcomers to the engineering disciplinary discourse community (Hyland 2004, 2015) on a Likert scale from 1 (disagree) to 6 (agree). The highest ranked formulas were compiled into a final list of 99 formulas and categorized according to their discursive function: referential expressions (e.g., at room temperature), stance expressions (e.g., assumed to be) and discourse organizing expressions (e.g., results indicate that) (Biber et al 2004). A correlation analysis reveals associations between the highest ranked formulas, their frequency in the corpus and their mutual information scores.
These findings contribute to engineering-specific writing instruction and learning by providing a list of pedagogically useful formulas. Further, they provide a contribution to the English for Specific Purposes movement with a methodology that can easily be replicated to create lists of other discipline-specific vocabulary. We conclude this report with pedagogical recommendations and future research directions.

Key words: *formulaic engineering language, corpus linguistics, English for Specific Purposes*

1. INTRODUCTION

This study was inspired by the creation of the Academic Formulas List (AFL; Ellis, Simpson-Vlach & Maynard 2008; Ellis & Simpson-Vlach 2009; Simpson-Vlach & Ellis 2010). The AFL includes academic formulas that are common in both speaking and writing, in addition to formulas that belong specifically to spoken and written domains. It is further categorized by discursive function. The creation of the list was accomplished using a formula teaching worth (FTW) equation, which was derived by triangulating corpus metrics with educational insights and psycholinguistic measures. Our aim was to partially replicate the design used to create the AFL in order to create an academic formulas list specific to written language in the discipline of engineering. Particularly, the corpus-driven approach used to identify formulaic language in the AFL was used and triangulated with pedagogical insights from experts from the engineering field.

2. LITERATURE REVIEW

2.1. Discipline-specific language

In order for newcomers to a field to participate in a given discipline-specific discourse community (Flowerdew 2002, Hyland 2015) such as the community of engineers, they must first be exposed to and acquire the language used by that community. Bhatia (1999) states that the acquisition of discipline-specific written conventions requires an awareness of the discursive procedures and practices, a "learning [of] the rules of the game" (Bhatia 1999, 26), before students are able to integrate the forms, functions, and social contexts of future professional communities (Tardy 2009). Such an awareness or "consciousness-raising" (Tardy 2009, 7), can be accomplished through explicit vocabulary learning (Li & Schmitt 2009).

One approach to creating an explicit awareness of the language used in a community is to identify the multi-word expressions employed in the discourse. Hyland (2008) recommends consulting published research articles as a source for the most commonly used language in a discipline. This can be accomplished using corpus linguistics tools. A small number of studies have used corpus linguistic approaches to investigate engineering texts (Luzón 2009, Ward 2007). Luzon (2009) investigated the use of authorial personal pronouns between engineering student learner corpora and "expert" technical writing field corpora. He found that learners tended to use *we* with less precision and rhetorical accuracy compared to the experts. Luzon proposes that future studies can incorporate genre analysis, expert corpora, and learner corpora; this combination could be a powerful pedagogical tool to assist students in raising their awareness of their language choices, the phraseology specific to their field of study, and specific patterns of rhetorical strategies.

In another study, ward (2007) utilized corpus linguistic tools to analyze the key vocabulary of engineering textbooks in order to offer pedagogical suggestions to those teaching english for specific purposes in foreign language contexts. his findings indicated that engineering texts are characterized by the formation of noun phrases (e.g., *gas phase reaction, gas temperature).* the present study aims to expand this line of inquiry by investigating technical writing in engineering and the use of corpus linguistics to inform the pedagogy thereof.

2.2. Formulaic language

Just as academic vocabulary has come to be regarded as important for language learning and testing (Coxhead 2000, Nation 2001), formulas (also known as multiword phrases, n-grams, lexical bundles, or chunks) are important units of language for the acquisition and use of academic language (Granger 1998, Li & Schmitt 2009). Formulaic language has been defined as "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is stored and retrieved whole from the memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002, 465). In other words, the building blocks of language may not be individual words, but sequences of words, which may have important implications for language learning and teaching.

In order for language learners and teachers to be able to take advantage of this formulaic organization of language, reliable measures are needed to identify formulas that will be useful. Insights from corpus-driven research, where an inductive approach is used to see which patterns emerge from a corpus (Biber 2009), have been informative in describing language use. However, Biber, Conrad and Cortes (2004) point out that corpus linguists have not come to a consensus on the defining characteristics and most appropriate methods for identifying formulaic language. Expanding on the work of Biber et al. (1999), they propose the identification of *lexical bundles*, or the "most frequent recurrent sequences of words" (Biber et al. 1999, 373) in a corpus of target language. This metric is easy to apply but produces long lists of frequent expressions that are not always useful.

One way to limit these lists is to set a frequency threshold such as 20 occurrences per million words (Biber, Conrad, & Cortes 2004) or to set a fixed expression length, such as 3-, 4-, or 5-grams. In addition, considering formulas that only occur in a wide range of genres or publications as was done in the creation of the Academic Word List (Coxhead 2000) can further reduce the size of unmanageable lists of frequent formulas. Simpson-Vlach and Ellis (2010) highlight that frequency is not a sufficient measure for identifying useful formulas, as some formulaic language is not particularly frequent (e.g., *longitude and latitude*) and some very frequent language is not intuitively formulaic (e.g., *and of the*). In addition to the formula's frequency, mutual information is another measure to consider. Mutual information is the strength of association a lexical bundle has within its words. For instance, *blue moon* has a high degree of mutual information while *red moon* may have little to none. Previous psycholinguistic experimental data and pedagogical expert opinions provide evidence for the importance of mutual information in developing the academic formulas list (Ellis, Simpson-Vlach, & Maynard 2008).

3. RESEARCH QUESTIONS

The following research questions guided our study:

1. What are the most frequently used lexical bundles in engineering writing (as opposed to other academic genres)?
2. Which formulas are most pedagogically useful for novice engineering students?

4. METHODOLOGY

4.1. Participants and context

The instructors of a first year undergraduate engineering course with a strong focus on writing participated in the current study by sharing their intuitions about formulaic engineering language on a survey. The impetus to create an Engineering Academic Formulas List (EAFL) came from working with the instructor and teaching assistants of this course in which approximately 700 students participate each semester. Several writing supports had been put into place, including a technical writing guide and memo templates created by the engineering faculty and a writing lab run by English Language Center writing experts. However, in spite of this support, the instructors noted that students, both domestic and international, were not prepared to enter the written engineering discourse community, which emphasizes objective writing using specific engineering terminology.

4.2. Corpora

4.2.1. Target corpora

The target corpora included approximately 1 million words of published engineering articles. The seed of this corpus were the published engineering texts (196,533 words) found in Hyland's (2004) research article corpus. The same 20 journals where these articles were sourced were then consulted and a selection of the most recent articles from each publication was compiled. These articles were converted into text files and cleaned to (roughly) match the composition of the original corpus. 897,335 words were added to increase the corpus to a total size of 1,093,868 words. These word counts are summarized in Table 1.

4.2.2. Comparison corpora

The remainder of the research articles from Hyland's (2004) corpus (868,636 words) plus a subset of written academic texts (690,438 words) from the British National Corpus (BNC; BNC Consortium, 2006) were compiled to form the comparison corpus (1,559,074 words).

Table 1 Word counts by source for the target and comparison corpora.

Target corpora: Engineering articles		Comparison corpora: Academic articles	
Hyland	196,533 words	Hyland academic	868,636 words
TigFox	897,335 words	BNC academic	690,438 words
Total	1,093,868 words	Total	1,559,074 words

4.3. Materials and procedure

The target and comparison corpora were entered into the *Collocate* (Barlow, 2004) program to generate 3-, 4- and 5-grams with their corresponding frequency and mutual information (MI) data. We selected n-grams that occurred at a rate of 20 occurrences per million in the engineering and general academic articles; the total number of lexical bundles in these two corpora were 2,250 and 2,489 respectively.

Next, the engineering and academic lists of lexical bundles were aligned to find the overlapping occurrences, and a log-likelihood calculation was used to identify which lexical bundles occurred significantly more frequently in the engineering corpus compared to the general academic corpus (Simpson-Vlach & Ellis 2010). Of the total number, approximately 1200 lexical bundles were significantly more frequent ($p<0.01$). This list was sorted first by frequency and three bands (high, mid, low) were created; three bands were similarly created for MI. Table 2 below contains sample lexical bundles that represent all variables: n-gram length (3, 4, 5), Frequency band (High, Medium, and Low; means 83.15, 31.90, and 23.57 per million respectively), and MI band (High, Medium, and Low; means 18.83, 11.27, and 8.08 respectively).

Following the identification of 1003 lexical bundles using the log-likelihood calculation, we further pared down the total to include bundles that occurred across a wide range of publications. We excluded any bundles that occurred in less than ten different publications, which resulted in 765 lexical bundles.

The next step toward to creating the EAFL was to gain engineering instructors' insights into which lexical bundles would be pedagogically useful to their students by distributing a survey. The 765 lexical bundles were divided into three groups and sent to twelve EGR 100 teaching assistants to rank each bundle according to the statement "this phrase is worth teaching in the EGR 100 course". The respondents ranked each item on a

Likert scale, from 1 (strongly disagree) to 6 (strongly agree). Four teaching assistants completed each survey. See Appendix A for a sample of the survey.

Table 2 Sample Lexical Bundles in Varying Frequency and Mutual Information Bands

Frequency band	MI Band (mean)		
(n per million)	Low (8.08)	Mid (11.27)	High (18.83)
	Is obtained by	By means of a	Make sure that
Low (23.57)	Limited by the	Have been proposed	Good agreement with the
	A decrease in	It is evident	Taking into account
	Decrease in the	Can be estimated	This is due to
Mid (31.90)	Variations in the	Can be derived	Be attributed to the
	Effects on the	Can be described	Results show that the
	Increase in the	Depending on the	Is shown in figure
High (83.15)	Indicates that the	In order to	It should be noted
	Note that the	Is assumed that	At room temperature

Inter-rater reliability between the three surveys was calculated. Survey A had 4 raters evaluate 255 items, inter-rater $\alpha = 0.46$. Survey B had 4 raters evaluated 255 items, inter-rater $\alpha = 0.72$. Survey C had 255 items and was rated by 4 raters, inter-rater $\alpha = 0.67$. A number of factors may account for the low inter-rater reliability in survey A. First, some of the raters had far more experience than others and would therefore have been more familiar with written engineering language. The raters also belonged to a variety of subfield specializations, which may mean that they had not been exposed to as wide a range of engineering formulaic language than they encountered in the survey.

5. RESULTS

One hundred eight lexical bundles received a score of 3.5 out of 6 or higher and were selected as the phrases with the highest teaching worth. This list was further scaled down by collapsing nearly identical phrases (e.g., *results indicate* and *results indicate that*) to bring the total number of phrases worth teaching to 99. (See Appendix B for the list). Following Biber et al. (2009) and Ellis et al. (2010)'s conventions, we categorized these 99 lexical bundles into the functional categories of referential (e.g., *at room temperature)*, discourse organizing expressions (e.g., *results indicate that*), and stance (e.g., *assumed to be)*. See Appendix C for categorized phrases.

6. DISCUSSION & CONCLUSION

Our first research question regarding the most frequently used lexical bundles in engineering writing (as opposed to other academic genres) resulted in approximately 765 lexical bundles identified. Our second research question and ultimate goal in the study was to determine the most pedagogically useful formulas, which yielded a final list of 99 engineering formulas. Given the difficulty that newcomers to a discourse community have in acquiring formulaic language, and the observed benefits of explicit learning of formulas (Li & Schmitt 2009), this list provides a contribution to instructors and learners in the field of engineering that can be used to acquire discipline-specific vocabulary.

A couple of limitations must be acknowledged and taken into consideration for future research. First, we did not obtain high reliability on Survey 1, and possible factors were explained above. Another limitation to our study is that graduate teaching assistants who are not yet fully-fledged published engineers themselves were assigned the role of expert raters. Finally, the wording of the questionnaire may have influenced the results. Some bundles may be considered more important for higher level engineering courses than EGR 100, so although they were scored as not worth teaching for first year students, they may still be useful to learn in more advanced years.

There is a lot of promise for future research in this vein. First, correlation analysis will be run on the list of 99 lexical bundles to determine any unifying characteristics (i.e., frequency, MI, and/or range) correlate most highly with teaching worth. We also hope to expand the exploration of technical engineering writing development to include information from novice writers in the first year technical writing course as well as data from the Michigan Corpus of Upper-level Student Papers (MICUSP). Finally, coordination with university-level EAP instructors will give us critical insight into the pedagogical integration of the Engineering Academic Formulas List into the classroom. Results from corpus linguistics can inform and transform evidence-based teaching practices across the curriculum. The Engineering Academic Formulas List is only the beginning to uncovering more discipline-specific discourse and pedagogy.

REFERENCES

Barlow, Michael. 2004. “Collocate”. *Houston: Athelstan Publications*.

Bhatia, Vijay K. 1999. "Integrating Products, Processes, Purposes and Participants in Professional Writing." *Writing: Texts, Processes and Practices*: 21-39.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman Grammar of Spoken and Written English*. Vol. 2. MIT Press.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. “*If You Look at…*: Lexical Bundles in University Teaching and Textbooks.” *Applied Linguistics.* Vol. 25, Issue 3: 371-405.

BNC Consortium. 2006. http://www.natcorp.ox.ac.uk/.

Coxhead, Averil. 2000. "A New Academic Word List." *TESOL Quarterly*: 213-238.

Dalton, David F., and David Dalton. 2008. "The foreign language engineering writer: What makes a readable memo report?". *The Asian ESP Journal* 4.

Ellis, Nick C., and Rita Simpson-Vlach. 2009. "Formulaic Language in Native Speakers: Triangulating Psycholinguistics, Corpus Linguistics, and Education." *Corpus Linguistics and Linguistic Theory* 5, no. 1: 61-78.

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. "Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus linguistics, and TESOL." *TESOL Quarterly* 42, no. 3: 375-396.

Fernandez-Parra, Maria, and Maria Leedham. 2013. "Writing in Engineering: Pronoun Usage in Written Assignments by Chinese, British and Greek Students."

Flowerdew, Lynne. 2002. "Corpus-based Analyses in EAP". *Academic Discourse:* 95-114.

Gimenez, Julio, and Juliet Thondhlana. 2012. "Collaborative Writing in Engineering: Perspectives from Research and Implications for Undergraduate Education. *“European Journal of Engineering Education* 37, no. 5: 471-487.

Granger, Sylviane. 1998. "Prefabricated Patterns in Advanced ESL Writing: Collocations and Formulae." In *Phraseology: Theory, Analysis and Applications,* edited by A. P. Cowie, 145-160. Oxford: Clarendon Press.

Hyland, Ken. 2002. "Options of Identity in Academic Writing." *ELT Journal* 56, no. 4: 351-358.

Hyland, Ken. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. University of Michigan Press.

Hyland, Ken. 2008. "As Can Be Seen: Lexical Bundles and Disciplinary Variation." *English for Specific Purposes,* no. 27: 4-21.

Hyland, Ken. 2015. *Academic Written English*. Shanghai: Shanghai Foreign Language Education Press.

Jenkins, Susan, Mary Kaye Jordan, and Patricia O. Weiland. 1993. "The role of writing in graduate engineering education: A survey of faculty beliefs and practices. *"English for Specific Purposes* 12, no. 1: 51-67.

Li, Jie, and Norbert Schmitt. 2009. "The Acquisition of Lexical Phrases in Academic Writing: A Longitudinal Case Study." *Journal of Second Language Writing* 18: 85-102.

McKenna, Bernard. 1997. "How Engineers Write: An Empirical Study of Engineering Report Writing." *Applied Linguistics* 18, no. 2: 189-211.

Molle, Daniella, and Paul Prior. 2008. "Multimodal Genre Systems in EAP Writing Pedagogy: Reflecting on a Needs Analysis." *TESOL Quarterly* 42, no. 4: 541-566.

Nation, Ian SP. 2001. *Learning Vocabulary in Another Language*. Ernst Klett Sprachen.

Saenkhum, Tanita. 2007. *Transfer of Knowledge from First-year ESL Writing Classes to Writing in the Disciplines: Case Studies of Writing Across the Curriculum*. Proquest.

Simpson-Vlach, Rita, and Nick Ellis. 2010. An Academic Formulas List (AFL). *Applied Linguistics, 31*, 487-512.

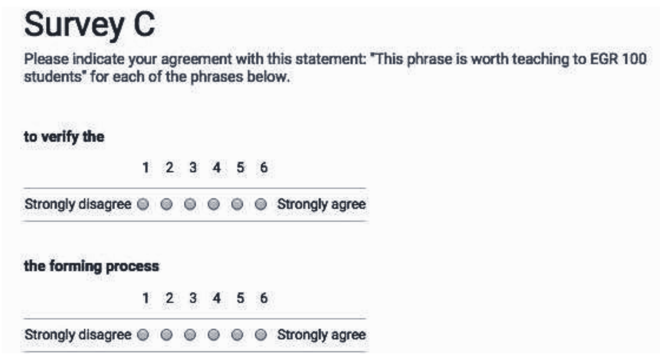
Smith, Summer. 2003. "What is "Good" Technical Communication? A Comparison of the Standards of Writing and Engineering Instructors." *Technical Communication Quarterly* 12, no. 1: 7-24.

Tardy, Christine. 2009. *Building Genre Knowledge*. West Lafayette, IN: Parlor Press.

Ward, Jeremy. 2007. "Collocation and Technicality in EAP Engineering." *Journal of English for Academic Purposes* 6, no. 1: 18-35.

Wheeler, Edward, and Robert L. McDonald. 2000. "Writing in Engineering Courses." *Journal of Engineering Education* 89, no. 4: 481-486.

APPENDIX A
Teaching Assistant Survey



Appendix B

Formulas worth teaching (according to survey results)

Lexical bundle	Survey score (out of 6)	Frequency	Mutual information
(are) shown in figure	3.50	73	6.92
a function of time	3.50	21	6.98
a positive effect on	3.50	25	6.99
a reduction in	3.50	21	7.2
a significant effect on	3.75	24	7.39
according to the	3.75	22	7.75
an important role	3.50	24	7.76
an increase in	4.00	25	7.86
are discussed in	3.50	26	7.86
are shown in table	3.75	22	7.87
are summarized in (table)	3.50	22	7.9
as a function of (time)	3.75	39	8.01
as described in	3.50	92	8.11
as indicated in	4.00	23	8.27
as mentioned in	3.50	31	8.27
as shown in figure	3.50	40	8.29
as shown in table	3.75	25	8.42
assumed to be	3.75	193	8.45
at room temperature	3.50	34	8.51
based on the	3.50	55	8.51
be attributed to the	3.50	61	8.53
be calculated by	4.25	66	8.54
be noted that the	3.75	63	8.66
be seen from	3.50	28	8.67
by a factor of	3.50	63	8.75
can be derived	3.75	32	9.1
can be described	3.75	31	9.16
can be determined	4.00	50	9.32
can be expressed as	4.33	23	9.33
can be observed	3.50	35	9.46
determined by the	4.50	34	9.47
diagram of the	3.75	78	9.51
distance from the	3.50	23	9.54
due to the fact that	3.75	38	9.61
equation of motion	3.50	44	9.62
focused on the	4.50	398	9.69
for a given	3.50	33	9.85
for different values	4.00	20	9.96
for each test	4.25	22	10.15
given in table	3.50	26	10.15
has a positive effect on	4.25	29	10.27
have been proposed	3.50	24	10.34
illustrated in figure	3.50	56	10.34
in order to (achieve)	3.50	23	10.48
in the next section	3.75	22	10.62
in this design	3.50	40	10.71
indicate that the	3.75	20	10.86
is assumed that	3.50	26	10.94

is considered to	3.75	24	10.98
is defined as the	3.75	25	11.12
is described in	4.00	38	11.15
is due to	3.50	24	11.37
is expected to	4.25	32	11.37
is expressed as	3.50	85	11.61
is illustrated in	4.50	132	11.69
is obtained by	4.33	24	11.82
is presented in	4.00	30	11.83
is proportional to the	4.25	26	11.9
is represented by	3.75	27	11.95
is required to	3.50	47	12.08
is shown in (figure, table)	4.00	25	12.09
is similar to	4.00	28	12.1
it is evident	3.50	32	12.11
it is observed that	3.50	28	12.21
it should be noted	4.50	46	12.44
limited by the	3.50	46	12.5
listed in table	3.50	195	12.71
noted that the	3.50	22	12.82
number of samples	3.75	27	12.88
parameters such as	3.75	21	12.95
plotted in fig	3.75	36	13.36
presented in figure	4.25	71	13.4
presented in table	3.50	26	13.49
presented in this	3.50	29	13.54
proportional to the	3.75	31	14.09
result in a	4.25	22	14.15
results and discussion	4.25	51	14.42
results are shown	3.50	568	14.55
results indicate that	4.25	42	14.63
results obtained from	3.75	20	14.78
seen from the	4.25	228	15.27
shown in table	4.25	29	15.32
significant effect on	4.00	71	16.47
summarized in table	3.50	28	16.65
the experimental data	3.75	31	16.99
the experimental results	3.50	27	17.04
the flow rate	4.50	44	17.22
the proposed method	4.75	43	17.36
the relationships between	4.50	352	17.46
the results show (that)	3.75	30	18.08
this indicates that	4.25	178	18.27
to calculate the	4.25	752	19.18
to describe the	3.75	25	19.54
to examine the	3.50	24	20.3
to illustrate the	4.00	71	21.29
to satisfy the	4.25	28	21.33
was found that	3.50	28	24.21
was found to	4.25	40	25.42
with respect to	3.50	24	29.35

Appendix C

Functional Categorization (based on Biber, Conrad and Cortes, 2004)

Referential expressions	Discourse organizing expressions	Stance expressions
▪ direct reference to physical or abstract entities, or the textual context itself	▪ relationships between prior and coming discourse	▪ attitudes or assessments of certainty that frame some other proposition
a function of time	according to the	an important role
a positive effect on	are discussed in	assumed to be
a reduction in	(are) shown in figure	be attributed to the
a significant effect on	are shown in table	can be derived
an increase in	are summarized in (table)	can be described
as a function of (time)	as described in	can be determined
at room temperature	as indicated in	can be expressed as
by a factor of	as mentioned in	can be observed
diagram of the	as shown in figure	determined by the
distance from the	as shown in table	due to the fact that
equation of motion	based on the	indicate that the
for a given	be calculated by	is assumed that
for different values	be noted that the	is considered to
for each test	be seen from	is expected to
in this design	focused on the	is required to
is defined as the	given in table	it is evident
is obtained by	has a positive effect on	it is observed that
is proportional to the	have been proposed	it should be noted
is similar to	in order to (achieve)	noted that the
limited by the	illustrated in figure	
number of samples	in the next section	
parameters such as	is described in	
proportional to the	is due to	
results and discussion	is expressed as	
results obtained from	is illustrated in	
the experimental data	is presented in	
the experimental results	is represented by	
the flow rate	is shown in (figure, table)	
the proposed method	listed in table	
the relationships between	plotted in fig	
the results show (that)	presented in figure	
with respect to	presented in table	
	presented in this	
	result in a	
	results are shown	
	results indicate that	
	seen from the	
	shown in table	
	significant effect on	
	summarized in table	
	this indicates that	
	to calculate the	
	to describe the	
	to examine the	
	to illustrate the	
	to satisfy the	
	was found that	
	was found to	