Computations of a frequently used vocabulary in technical documents to help students reading English technical textbooks in technological and vocational college/university programmes

Bing-Yuh Lu, Ming-Li Tung, Jo-Cheng Hung, Yu-Szu Lin, Yuan-Hsin Huang, Chen-Hui Chung, Yung-Cheng Chien & Kung-Ming Chang

Tung-Nan Institute of Technology
Taipei, Taiwan

ABSTRACT: The vocabulary of two popular textbooks on computer architecture and organisation and the datasheets of five popular processors are counted in order of repeated times. The textbooks' authors are M.M. Mano and W. Stallings. The five microprocessors are Intel's Pentium, NEC's ARM7TDMI, IBM's PowerPC 405GPr Embedded Processor, Philips' 80C51/87C51/80C52/87C52 and Texas Instruments' Digital Signal Processor TMS320VC5441. The top 500 ordered vocabulary comprises over 80% of the total 125,420 words in Mano's book (repetitive rate: 102,442/125,420). The top 700 repetitive vocabulary comprises over 80% of the total 109,553 words in Stallings' textbook (repetitive rate: 89,440/109,533). The top 800 ordered vocabulary comprises over 80% of the total 41,485 words in the aforementioned datasheets of the five popular microprocessors (presented rate: 33,343/41,485). Compared with the frequently used vocabulary of these two textbooks, more than 80% of the same vocabulary is presented in the top 1,300 vocabulary of both textbooks (presented rates are Mano: 104,207/125,420, Stallings: 87,848/109,553), ie if a student only remembers a vocabulary of 929 English words, then he/she can read about 80% of the words in both textbooks.

INTRODUCTION

It is difficult to teach with technical English textbooks in non-native English speaking nations because people in these countries use their mother tongues in their daily life. English is very important because Western technology and social science influence and impact on non-native English speaking nations. Furthermore, English has become the official language of the Internet [1].

The applications of rapidly developing information technologies (ITs) on the Internet are the dominant knowledge sources in developed and developing countries; as such, the English proficiency level of a country will determine its development of a knowledge-based economy, which is the focus of future economic boom [1-4]. Technological and vocational college/university (TVCU) programmes in non-native English speaking nations play important roles in the implementation of vocational technologies. Therefore, technical English is very important for TVCU programmes.

Psychologically speaking, when students in non-native English speaking nations first have to read technical English textbooks, most of them feel a lot of pressure. As such, these students try to avoid reading them, especially if there are native translated versions of the same textbooks. This is because students not only have to know the vocabulary, phrases and sentence structures in English technical textbooks, but also have to understand the meanings of the principles and technologies in the textbooks.

Students have to spend much more time reading the English technical textbooks without the native translated versions. This is why, when teachers usually assign English technology books as textbooks, students still seek to find the native translated versions of these textbooks in order to study these courses. It is easier for students to absorb the new knowledge through reading the native versions of the technical textbooks.

In fact, it was recently shown in Taiwan that the salary of employees is dependent on their English abilities [5]. Proficiency in the English language is considered to be a passport to obtain a better job [6]. This means that the better a person's English skills are, the higher his/her salary. This phenomenon is global, and shows how English abilities play an important role in TVCU programmes.

The purpose of this study is to compute the frequently used vocabulary in the field of computer architecture and organisation. The vocabulary of two popular textbooks about computer architecture and organisation and the datasheets of five popular processors have been counted with regard to the presented vocabulary and allocated sequentially accumulative percentages. The authors of the textbooks are Prof. M.M. Mano, and Prof. W. Stallings, respectively [7][8].

The five microprocessors are as follows:

- Intel's Pentium;
- NEC's ARM7TDMI;
- IBM's PowerPC 405GPr Embedded Processor;
- Philips' 80C51/87C51/80C52/87C52;
- Texas Instruments' Digital Signal Processor TMS320VC5441.

Other researchers have proposed the concept of a vocabulary spectrum [9-11]. Their results can verify this concept and help students in non-native English speaking nations overcome psychological barriers in reading the technical English textbooks so as to elevate their English abilities.

METHODS

The vocabulary of two textbooks and the datasheets of the five microprocessors have been computed in this study. They have to be scanned page by page to the BMP file, and be transferred into Microsoft *Word* format by character recognition software. There are some errors in the process of the character recognition. They are the number "0", letters "O", "o" and "Q", number "1", letters "l", "I" and "L", letters "p", "q" and "o", letters "m" and "n", and letters "P", "D" and "B". These have to be checked and corrected manually.

The corrected page was then duplicated as a long string in an input box programmed using Microsoft *Visual Basic* (VB). This program read the string array letter by letter, and added characters in a temporary string variable (Word_Temp). When the program read the character of ".", ",", " ", "(", ")", ":", "[", "]", "?", """, it compared the Word_Temp with every element in the array (Comp_Array). This array saved the word presented in the page. If the Word_Temp hit in the Comp_Array, then the repetitive time of the word was increased, or else the new word was added into the Comp_Array. The Comp_Array was saved to compare the vocabulary of the next page. The flowchart of this program is shown in Figure 1.

After having finished the computation of the presented times of one of the textbooks, the Comp_Array was saved as a Microsoft *Excel* file, before resetting the Comp_Array and continuing to count the next textbook. The datasheets of the five microprocessors were put together and computed as a textbook. From this, one could obtain the repeated times of vocabulary within each textbook (the datasheets were put together as a textbook) and determine the total words of each textbook.

Finally, the vocabulary of the textbooks was chosen in order of repetitive times. The accumulated vocabulary was 1,800 with a step of 100, with a repetitive percentage of the total number of words accumulated from each textbook. The results present the frequency of the vocabulary in each textbook.

By coding the Visual Basic Applications (VBA) in Microsoft *Excel*, the overlapped vocabulary can be compared with the datasheets of the five microprocessors. The repeated times of the overlapped vocabulary from top one to 1,800 with steps of 100 were computed, along with the accumulated percentage of the total number of the same words in each textbook. The results indicate the most frequently used vocabulary between these two textbooks.

RESULTS

The totals of the vocabulary and words of the two textbooks and the datasheets are shown in Table 1. The textbook written by Mano has been accounted for in order to get the following totals: the number of words is 125,420, while vocabulary stands at 5,150. The totals of words and vocabulary are 109,553 and 6,694, respectively, in the textbook written by Stallings. The same computation results are 41,485 and 4,974, respectively, in the datasheets of the five microprocessors. The results indicate that some vocabulary is highly repeated in the technical documents.

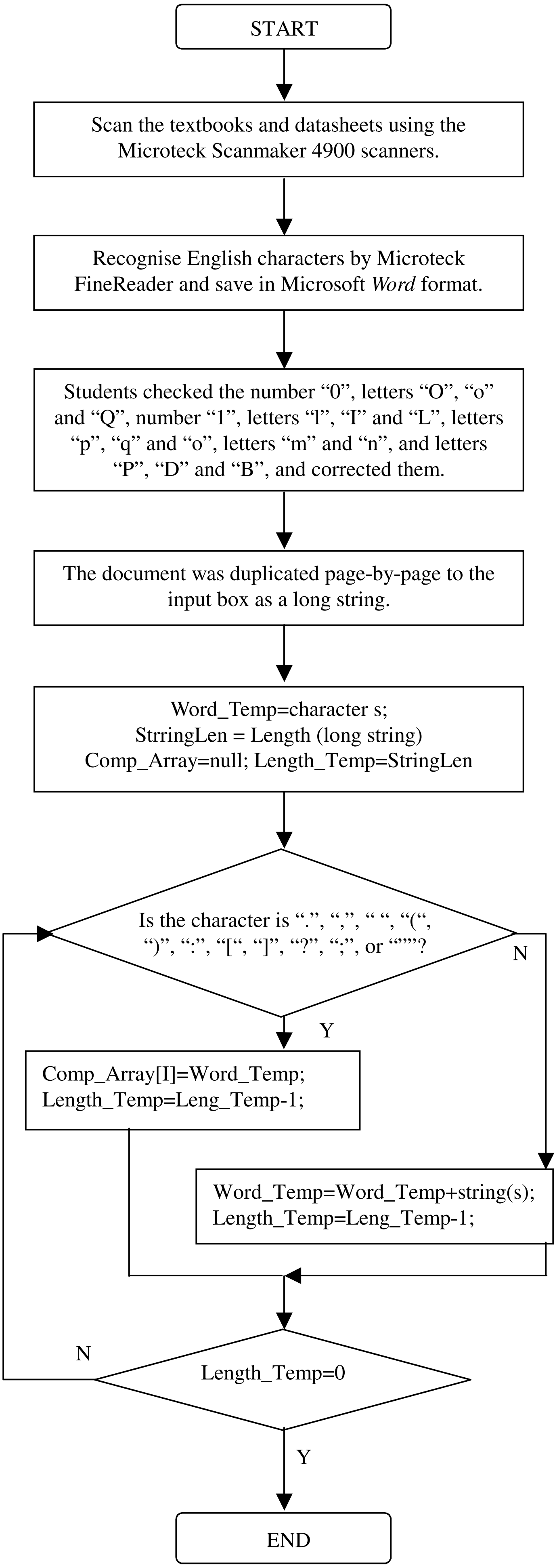
START
Scan the textbooks and datasheets using the Microteck Scanmaker 4900 scanners.
Recognise English characters by Microteck FineReader and save in Microsoft *Word* format.
Students checked the number "0", letters "O", "o" and "Q", number "1", letters "l", "I" and "L", letters "p", "q" and "o", letters "m" and "n", and letters "P", "D" and "B", and corrected them.
The document was duplicated page-by-page to the input box as a long string.
Word_Temp=character s;
StrringLen = Length (long string)
Comp_Array=null; Length_Temp=StringLen
Is the character is ".", ",", " ", "(", ")", ":", "[", "]", "?", ";", or """?
N
Y
Comp_Array[I]=Word_Temp;
Length_Temp=Leng_Temp-1;
Word_Temp=Word_Temp+string(s);
Length_Temp=Leng_Temp-1;
N
Length_Temp=0
Y
END

Figure 1: The flowchart of the vocabulary computation program.

Table 1: Total vocabulary and words of each of the textbooks and the datasheets.

	Mano	Stallings	Datasheets
Total vocabulary	5,150	6,694	4974
Total words	125,420	109,553	41485

The results for vocabulary computations of the textbooks and datasheets are shown in Figure 2. The x-axis represents the vocabulary by repeated times with steps of 100, while the y-axis represents the accumulated percentage of the presented times of the chosen vocabulary. If students only remember the top 500 elected vocabulary of the textbook written by Mano, then they can read 81.7% of the words in this textbook. Alternatively, students who remember only the top 700 sorted vocabulary of the textbook written by Stallings can read 81.6% of the words in this textbook. Furthermore, students who remember only the top 800 sorted vocabulary of the datasheets of the five microprocessors can read 80.4% words in this textbook. If students can remember the top 1,800 sorted vocabulary of the textbooks and datasheets, then they can read over 90% of the words in all of them.

From the aforementioned results, the overlapped vocabularies of these two textbooks were selected for further study. These textbooks present the same topics, especially the basic concept of computer architecture. Therefore, some technical vocabulary and terminologies were frequently used in both textbooks.

A comparison of the presented times of the same vocabulary between the two textbooks, with steps of 100, is shown in Table 2. Students can select either of the two textbooks, and if they remember the top 1,500 selected vocabulary, and given the score of 929 overlapped vocabulary, the percentages of the same vocabulary are 83.09% and 80.19%, respectively.

The reading of datasheets and user guides of the electronic devices are very important in jobs that are related to the field of electronic engineering. Will the activity of reading English technical textbooks help students to also read datasheets or user guides more easily? The answer to this can be found in comparisons of the vocabulary between datasheets and textbooks. As the results in Table 3 indicate, the repeated times of the same vocabulary between the textbook written by Mano and the five datasheets of the popular microprocessors were compared with the step of 100. The findings show that the accumulated percentage of the number of the same vocabulary decreased with the increasing chosen vocabulary by the presented times, but that the accumulated percentage of the repetitive times increased.

Students who read Mano's textbook and remembered the top 1,500 vocabulary covered the same 588 vocabulary of the datasheets of the five popular microprocessors, and the percentages of repeated times of the same vocabulary are 74.19% and 60.73%, respectively.

The same comparison for the textbooks written by Stallings is presented in Table 4. Students who read Stalling's textbook and remembered the top 1,500 vocabulary could cover the same 585 vocabulary in the datasheets of the five popular microprocessors. The percentages of the repetitive times of the same vocabulary for Stallings' textbook were found to be close with the results obtained for Mano's book.

DISCUSSIONS

As the results above show, the authors have finished translating English vocabulary into a Chinese table for the top 800 vocabulary sorted by the repetitive times for the textbooks, and asked students in the Department of Electronic Engineering at Tung-Nan Institute of Technology, Taipei, Taiwan, to memorise them at the beginning of the semester.

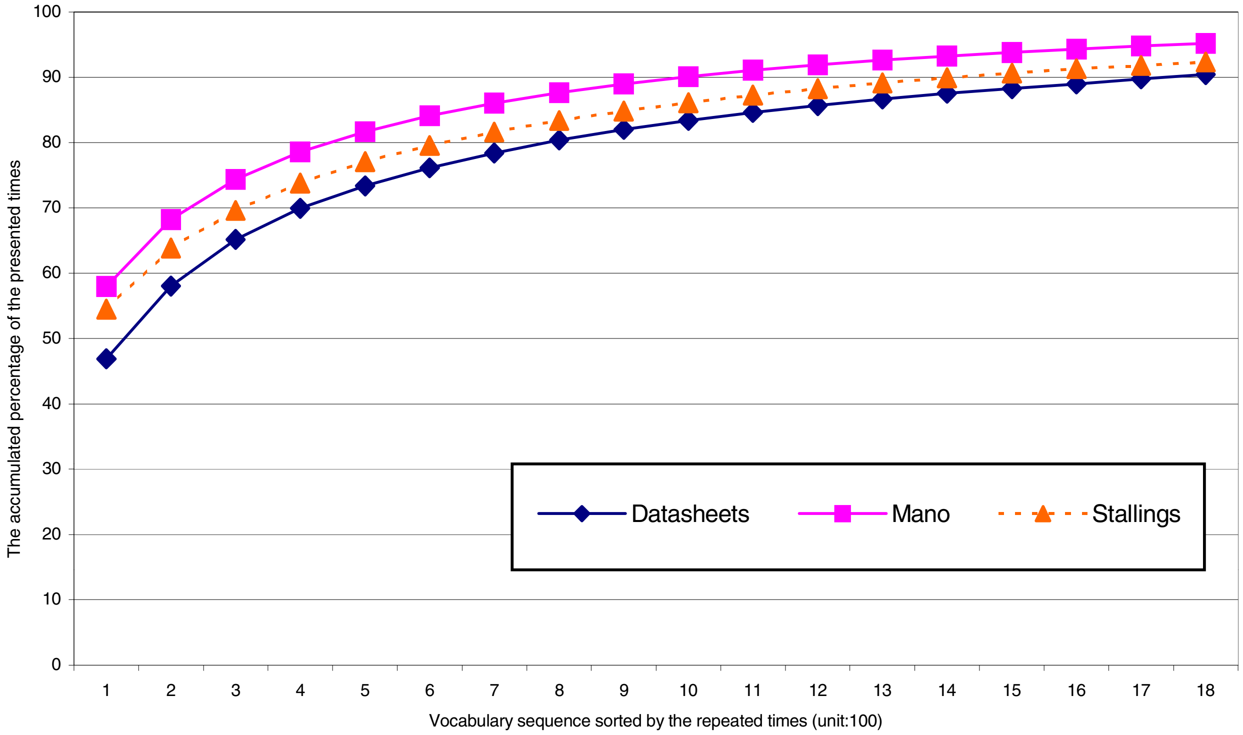

Figure 2: Vocabulary computation of the textbooks and datasheets.

Table 2: A comparison of the repeated times of overlapped vocabulary between the two textbooks with a step of 100 (the total words of each textbook are Mano: 125,420 and Stallings 109,553).

Vocabulary Sorted by the Repeated Times with a step of 100	Number of Overlapped Vocabulary		Repetitive Times of Overlapped Vocabulary (Mano)		Repetitive Times of Overlapped Vocabulary (Stallings)	
	Accumulated Number	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)
100	64	64.00	62,804	50.07	51,456	46.97
200	131	65.50	73,165	58.34	60,974	55.66
300	188	62.67	79,524	63.41	65,705	59.98
400	251	62.75	84,322	67.23	69,663	63.59
500	314	62.80	87,779	69.99	72,744	66.40
600	377	62.83	90,908	72.48	75,176	68.62
700	442	63.14	93,270	74.37	77,692	70.92
800	516	64.50	95,402	76.07	80,242	73.24
900	574	63.78	97,045	77.38	81,724	74.60
1,000	638	63.80	98,723	78.71	83,093	75.85
1,100	702	63.82	100,013	79.74	84,354	77.00
1,200	763	63.58	101,456	80.89	85,505	78.05
1,300	814	62.62	102,348	81.60	86,286	78.76
1,400	878	62.71	103,348	82.40	87,100	79.50
1,500	929	61.93	104,207	83.09	87,848	80.19
1,600	990	61.88	104,863	83.61	88,729	80.99
1,700	1,045	61.47	105,663	84.25	89,247	81.46
1,800	1,107	61.50	106,283	84.74	89,875	82.04

Table 3: A comparison of the repeated times of the overlapped vocabulary between the Mano's textbook and the five data sheets of the popular microprocessors with the step of 100 (the total words of each are Mano: 125420 and datasheets: 41485).

Vocabulary Sorted by the Repeated Times with the Step of 100	Number of the Overlapped Vocabulary		Repetitive Times of the Overlapped Vocabulary (Mano's)		Repetitive Times of the Overlapped Vocabulary (Stallings')	
	Accumulated Number	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)
100	49	49.00	56,796	45.28	12,645	30.48
200	85	42.50	65,788	52.45	15,162	36.55
300	125	41.67	71,462	56.98	17,399	41.94
400	160	40.00	74,682	59.55	18,646	44.95
500	210	42.00	77,888	62.10	19,939	48.06
600	249	41.50	80,705	64.35	20,633	49.74
700	288	41.14	83,082	66.24	21,297	51.34
800	330	41.25	84,433	67.32	22,103	53.28
900	369	41.00	85,466	68.14	22,830	55.03
1,000	417	41.70	87,068	69.42	23,659	57.03
1,100	457	41.55	89,019	70.98	24,194	58.32
1,200	487	40.58	90,047	71.80	24,407	58.83
1,300	516	39.69	91,176	72.70	24,574	59.24
1,400	551	39.36	92,455	73.72	24,945	60.13
1,500	588	39.20	93,051	74.19	25,193	60.73
1,600	632	39.50	93,964	74.92	25,450	61.35
1,700	677	39.82	95,190	75.90	25,725	62.01
1,800	699	38.83	95,559	76.19	25,877	62.38

The vocabulary in this table includes over 80% of the total words of the textbooks, and should help students to directly read each of the two English computer architecture textbooks utilised in this study without having to resort to using the Chinese versions. The feature of this table is that one word has only one or two Chinese meanings just for the textbook. Students should not be confused about the meaning of the vocabulary in the textbook. Although students usually look up a dictionary patiently for meanings, they still cannot be sure of the correct meaning.

Table 4: A comparison of the repeated times of the same vocabulary between Stallings' textbook and the five data sheets of the popular microprocessors with the step of 100 (the total words of each are Stallings: 109533 and datasheets: 41485).

Vocabulary Sorted by the Repeated Times with the Step of 100	Number of the Same Vocabulary		Repeated Times of the Overlapped Vocabulary (Mano)		Repeated Times of the Overlapped Vocabulary (Stallings)	
	Accumulated Number	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)	Accumulated Times	Accumulated Percentage (%)
100	44	44.00	45,918	41.91	10,435	25.15
200	80	40.00	53,574	48.90	12,519	30.18
300	116	38.67	58,088	53.02	13,831	33.34
400	157	39.25	60,960	55.64	15,320	36.93
500	210	42.00	63,865	58.30	16,237	39.14
600	236	39.33	66,194	60.42	17,097	41.21
700	272	38.86	68,311	62.35	17,654	42.56
800	315	39.38	69,687	63.61	18,137	43.72
900	351	39.00	70,715	64.55	18,557	44.73
1,000	391	39.10	71,997	65.72	19,208	46.30
1,100	432	39.27	73,389	66.99	19,510	47.03
1,200	463	38.58	74,336	67.85	19,866	47.89
1,300	507	39.00	75,183	68.63	20,288	48.90
1,400	552	39.43	76,708	70.02	20,723	49.95
1,500	585	39.00	77,265	70.53	21,117	50.90
1,600	631	39.44	78,141	71.33	21,660	52.21
1,700	680	40.00	79,318	72.40	21,927	52.86
1,800	710	39.44	79,605	72.66	22,127	53.34

Furthermore, the authors are going to design a computer program for specific textbooks on electronic engineering. When students key in a particular English word, the program will show the Chinese translation of this word, as well as the number of times that this word is repeated in the textbook. This should then help students to remember frequently utilised vocabulary in the textbook.

As the results have shown, the numbers of vocabulary in the textbooks are much less than the total tally of words. Students remember the top 1,500 chosen vocabulary by the number of times that they are presented in the textbooks and datasheets of the five popular microprocessors. As such, students can read most of the technical documents.

CONCLUSIONS

The following conclusions have been reached from this study:

- The results verify the concept of a *vocabulary spectrum* that may be different in various professional fields.
- The overlapped vocabulary between textbooks and relevant datasheets can help students to read technical documents more easily in their jobs, especially for those in non-native English speaking nations.
- It is very important to teach English technical textbooks in non-native English speaking nations so that there will be smoother transition for students into the workplace after graduation by facilitating this form of training.
- The results help in the design of curricula for both common English courses and technological professional courses in TVCU programmes.

ACKNOWLEDGEMENTS

The authors would like to thank for the support provided to the curriculum simulation project at Tung-Nan Institute of Technology, Taipei, Taiwan, for school system links in technological and vocational education in 2002, by the Ministry of Education, and National Science Council project number NSC 92-2218-E-236-001.

REFERENCES

1. Wolk, R.M., The effect of English dominance of the Internet and the digital divide. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 174-181 (2004).
2. Kjell, B. and Gailer, H., Enhancing usability by international students for a distance Web site. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 104-106 (2004).
3. Wilkens, E.J. and Brickman, M.S., Evolution of computer science program toward globalized technical education. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 139-143 (2004).
4. Chakrapani, P.N. and Ekbia, H.R., Opening up technological education: the perspective from social informatics. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 144-148 (2004).
5. http://tw.news.yahoo.com/2003/12/11/leisure/cna/4418116.html
6. Medgyes, P., *The Non-Native Teacher*. In: Holden, S. (Ed.), MEP Monographs. Hampshire: Macmillan Publishers, 1-8, (1994).
7. Mano, M.M., *Computer System Architecture* (3rd edn). Englewood Cliffs: Prentice Hall (1994).
8. Stallings, W., *Computer Organization and Architecture* (6th edn). Englewood Cliffs: Prentice Hall (2003).

9. Lu, B-Y., Tung, M-L., Wang, M-Y., Liu, H-L., and Shih, T-C., Systematically designed license exams in non-native countries to accelerate globalizing pace. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 9-16 (2004).
10. Tung, M-L., Lu, B-Y., Liu, H-L., Wang, M-Y., Luh, J-J. and Ju, K-C., English vocabulary spectrum analysis for the technological and vocational college/university programs in non-native English speaking nations. *Proc. IEEE Inter. Symp. on Technology and Society*, Worcester, USA, 96-101 (2004).
11. Lu, B-Y., Tung, M-L., Lin, Y-S., Huang, Y-H., Hung, J-C., Chung, C-H., Chang, K-M. and Chien, Y-C., Statistics of the vocabulary in the textbook for the course of computer organisation and architecture in the department of electric engineering, Tung-Nan Institute of Technology. *J. Electronic Engng. of TNIT*, 5, 17-26 (2004).