A Computer Science Word List

Daniel E Minshall

Submitted to Swansea University in fulfilment of the requirements for the Degree of Master of Arts (MA TEFL)
Swansea University, 2013

Summary:

This study investigated the technical vocabulary of computer science in order to create a Computer Science Word List (CSWL). The CSWL was intended as a pedagogical tool in the instruction of non-native English speakers who are studying computer science in UK universities. In order to create this technical word list, a corpus of 3,661,337 tokens was compiled from journal articles and conference proceedings covering the 10 sub-disciplines of computer science as defined by the Association for Computing Machinery (ACM). The CSWL was intended to be supplemental to both the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000) and was created using the criteria established by Coxhead (2000) for word selection. The CSWL contained 433 headwords and in combination with the GSL and AWL accounted for 95.11% of all tokens in the corpus. This was sufficient to meet the lexical threshold for sufficient understanding of a text as proposed by Laufer (1990). This study also conducted research into the technicality of the CSWL by comparison to other corpora, comparison to a technical dictionary and an investigation of the distribution of its headwords against the BNC frequency bands. Overall, the CSWL was found to be highly technical in nature. The final part of the research looked into the existence of multi-word units in computer science literature to build a Computer Science Multi-Word List (CSMWL) from the same corpus. A total of 23 items comprised the CSMWL and they were again chosen using the same criteria of range and frequency as established by Coxhead (2000). The CSMWL showed that whilst multi-word units do exist in computer science literature, they are mostly compound nouns with domain specific meaning.

DECLARATION:

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ……. (candidate)
Date …….

STATEMENT 1

This dissertation is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed …….. (candidate)
Date ……..

STATEMENT 2

I hereby give my consent for my dissertation, if relevant and accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signed ……. (candidate)

RECORD OF SUPERVISION

2012 - 13

NB: This sheet must be brought to each supervision and submitted with the completed Dissertation.

(The following record must be completed as appropriate by student and supervisor at the end of each supervision session, and initialed by both as being an accurate record. **NB it is the student's responsibility to arrange supervision sessions and he/she should bear in mind that staff will not be available at certain times in the summer**). If any of these supervisions are conducted by email, Skype or any other electronic means, this should be clearly indicated in the 'Notes' column.

Student Name: Daniel E Minshall
Student Number: 148047
Dissertation Title: A Computer Science Word List
Supervisor: Dr. Vivienne Rogers

Supervision	***Date, duration***	***Notes***	***Initials Supervisor***	***Initials student***
1: Brief outline of research question and preliminary title (**by pre June**)				
2: Discussion of detailed plan and bibliography (**by June**)				
3: Progress report, discussion of draft chapter (**by August**)				
4: (optional) progress report (**by September)**				
5: Submission **(September)**				

Statement of originality: ***I certify that this dissertation is my own work and that where the work of others has been used in support of arguments or discussion, full and appropriate acknowledgement has been made. I am aware of and understand the University's regulations on plagiarism and unfair practice as set out in the 'School of Arts and Humanities Handbook for MA Students', and accept that my dissertation may be copied, stored and used for the purposes of plagiarism detection.***

Signed:.....…… Date: ……..

Contents:

1. Introduction

Data from the Higher Education Statistics Agency (HESA) (http://www.hesa.ac.uk/) shows that 17.4% of all students studying in UK universities for the academic year 2011/2012 were non-UK domicile. This is an increase by 1.6% over the previous year. Also, for the same year, 4.68% of all students who graduated from UK universities did so in a computer science related subject. This suggests that as the number of non-native English speakers studying in UK universities is on the increase and as computer science is such an influential and widely studied discipline, there is an increasing demand for pedagogical tools to be designed for the assistance of L2 English learners studying this subject.

English for Specific Purposes (ESP) teaching and its sub-branches like English for Academic Purposes (EAP) have grown in recent years to accommodate for the increased intake of non-native speakers into UK universities (Jordan, 2002). This area of teaching has been assisted in particular by research into the vocabulary used in an academic English environment (e.g. Coxhead, 2000). One of the findings of this research is that whilst there may be considered a general academic vocabulary, it is also important to instruct students in the technical language they will utilise in their own studies. There is a difference in the language of medicine compared to engineering. This study intends to compensate for the lack of any research conducted into the specialist lexical needs of non-native speakers studying computer science through an English medium. Specifically it intends to compile a technical word list for computer science students derived from corpus analysis.

In order to accomplish this, it will first review the current state of vocabulary research (chapter 2). This is intended to provide the motivation for the research questions which will be proposed (chapter 3). Following this, there will be an outline of the methodology utilised in building the corpus data and extracting a word list from it (chapter 4), and the results of these will then be demonstrated (chapter 5). Once these results have been obtained they will be discussed in light of findings from other similar studies (chapter 6), before this paper is concluded (chapter 7) and any suggestions for further research are disclosed (chapter 8).

2. Literature Review

2.1 Introduction

The purpose of this study was to investigate the viability of a technical word list for the discipline of computer science. This Computer Science Word List (CSWL henceforth) was designed to act as a supplement to both the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000). The intention was that it would be possible for an L2 learner of English studying this subject in an English medium to use such a list as a pedagogical tool.

In order to accomplish this, it is first necessary to discuss the current state of research into vocabulary relevant to this study. This review will show how a word may be defined and categorised, the idea of frequency and the notion of vocabulary size, coverage and their relationship with comprehension. It will also consider the technicality of a word, the increasing importance in recent research on collocational behaviour and how phrasal expressions could be included in a supplemental word list. Finally, it will discuss studies which have created technical word lists in other disciplines as well as the GSL and AWL.

2.2 Definition and categorisation of words

One major research problem involved in studying vocabulary is that, as a unit of measurement, the word can be difficult to clearly define and hence count (Milton, 2009). Any empirical study, in particular a quantitative study such as this one, requires precisely countable units of measurement. However, there are many different ways of counting words and this can lead to huge variations in numbers, if these different methods are employed. The typical example given of this is the count made of word families in Webster's Third International Dictionary by Nation (2001a) and Schmitt (2000). Their counts were 54,000 and 114,000 respectively. This level of disagreement illustrates the need for clear criteria to be established in the way words may be counted.

A simple count of words in a text can be restricted to tokens and types. Tokens are the number of running words in a text whilst types are the unique occurrence of each word in any text. Using both of these measurements will commonly lead to different values for each because of the re-occurrence of common (often function) words as in the sentence *He ate his dinner before he went out*. There

are 8 tokens here, but only 7 types due to the repetition of the word *he*. Tokens are important for any corpora-based research (e.g. Coxhead, 2000; Konstantakis, 2010) as this is the unit by which the size of the corpora are counted. Types are relevant to a technical word list as they are the morphemes of a headword which are included in its expansion.

Words may have many lexical types so they become increasingly difficult to count once their inflections and derivations are considered. Inflections influence a word's grammatical properties, often for verb agreement purposes or plural meaning, but do not change the part of speech, such as with *quick* and *quickest*. Words counted this way are known as lemmas. Derivations are affixes to a base word which can change its word class and meaning, such as *quick* to *quickly*. Counting words in this way, including both derivations and inflections, is called using word families. Previous technical word list studies (e.g. Coxhead, 2000; Coxhead & Hirsch, 2007; Konstantakis, 2007; Wang et al, 2008) have employed word families to produce fully expanded versions of the headword entries in such lists.

The concept of word families means that words such as *visible*, *invisible* and *invisibility* are all considered part of the same word family. This greatly reduces the number of words in the English language, based on the assumption that there is little or no extra burden for learning the inflections and derivations of a word once the grammatical rules for doing so are known by a non-native speaker (Nation, 2001a). Research from other studies has shown that this is likely. Levelt (1989) proposed a model for how speech is produced, which was partly based on previous work on speech errors (Fromkin, 1973). It demonstrated how encoding of language is a function of the brain called formulation which is in turn dependent on a lexical store which must contain the various inflections and derivations of a word.

However, not all inflections and derivations of a word provide a similar learning burden due to irregularities and this may lead to overestimates of a student's vocabulary size (Laufer, 1990). For this reason, Nation and Bauer (1993) created a list of 7 levels across which word families could be organised, based on the frequency and regularity of their affixes. They considered level 7 words as those which have classical roots and affixes which need to be learnt separately by both native and non-native speakers alike. These word families contain a headword with an expanded list of types of the same headword. For example, the headword *act* contains the types *acted, acting* and *acts*, but not *actor* as the *-or* affix is not within level 6 expansion. With this as a limiting factor on how a lexical type may be considered part of a word family, Coxhead (2000) decided upon level 6

expansion for headwords within the AWL. Further research, which built on the AWL, has also agreed on this level of expansion (e.g. Coxhead & Hirsch, 2007; Konstantakis, 2007; Wang et al, 2008). Table 2.1 illustrates how a word may be expanded as far as level 6.

Level 1
A different form is a different word. Capitalization is ignored.
Level 2
Regularly inflected words are part of the same family. The inflectional categories are - plural; third person singular present tense; past tense; past participle; -ing; comparative; superlative; possessive.
Level 3
-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, all with restricted uses.
Level 4
-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-, all with restricted uses.
Level 5
-age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom; officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery; trickery), -ese (Japanese; officialese), -esque (picturesque), -ette (usherette; roomette), -hood (childhood), -i (Israeli), -ian (phonetician; Johnsonian), -ite (Paisleyite; also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory), -ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise; discussion-wise), anti- (anti-inflation), ante- (anteroom), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage; enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify; subterranean), un- (untie; unburden).
Level 6
-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-.

Table 2.1 Word family levels (Nation, 2012)

As has been seen, there are precise ways in which a word may be counted. Understanding the definition of tokens, types and word families (expanded to level 6) is central to any technical word list study.

2.3 Frequency, vocabulary size, coverage and comprehension

The frequency of a word is generally defined as the number of times it occurs (as a word family) per token of a text. Whilst it is still an assumption that a word's frequency has an effect on its ability to be learnt, it seems a reasonable one to make and there is strong supporting evidence for this claim. Milton (2006) obtained a statistically significant relationship between frequency bands and vocabulary size scores using an ANOVA ($F=93.727$, $p<0.001$) on a study of 227 L2 English learners in a Greek school. These frequency bands have been produced through frequency-based corpora research (e.g Nation & Heatley, 2002). They are a way of demonstrating how common a word is by grouping them together in 1,000 word bands, such that the 1k frequency band contains the most common 1,000 words in a language and the 2k frequency band contains the second most common 1,000 words and so on. If frequency suggests a word is more likely to be learnt, then the more frequent a word, the more likely it is to reoccur in a text and the more likely it is to be understood by the reader. This leads to the concept of coverage: the percentage of tokens in a text that an L2 reader understands.

Connected to the notion of coverage is the idea of a lexical threshold. This figure is also given as a percentage and represents the coverage required for sufficient comprehension of any given discourse. Laufer (1989) studied 100 L2 learners of English at the University of Haifa. They were given a reading comprehension exercise in which they were also asked to mark the vocabulary they understood. Those students who had lexical coverage of at least 95% of the text performed significantly better than those who did not, which allowed Laufer to postulate this figure as a minimal threshold for comprehension. However, Laufer's definition of significantly better performance was based on an unconventional evaluation: the minimum pass mark for an examination at the University of Haifa, which was set at 55%. Whilst this particular level of performance might be considered arbitrary, Laufer did at least demonstrate a significant effect at this point. Liu and Nation (1985) also agreed with this lexical threshold of 95%.

Hu and Nation (2000) tested 66 L2 learners of English, who studied at university level and had performed well on a Vocabulary Levels Test (Nation, 1983), with a reading activity. Such tests are an attempt to evaluate an L2 learners vocabulary size by sampling their understanding of words from increasingly less common frequency bands, which also contain dummy words set to control for guessing. Their results demonstrated a predictable relationship between comprehension and

unknown word density and found that whilst 90-95% coverage was sufficient for some of their subjects to perform adequately, far better results were obtained at 98% coverage. Their findings agreed with an earlier paper by Hirsch and Nation (1992), which investigated the relationship between lexical coverage and known and unknown words. They noticed a non-linear connection between these 2 factors such that there was a steep drop in word density known at the 98% mark. A more recent study (Schmitt, Jiang & Grabe, 2011) revealed a linear relationship between coverage and comprehension in the reading ability of 611 L2 learners of English. In effect, they found no threshold value indicated by a sudden increase in comprehension at any level of coverage, but again concluded that 98% coverage was a more reliable target.

The research to date would suggest that for optimal comprehension of any text there is a threshold of coverage between 95-98%. However, this is not an absolute value. It should be remembered that this is an ideal figure and that sufficient comprehension of a text with a lower threshold coverage may be obtained. The problem with setting the threshold to 98% is one of diminishing returns. There is not a linear relationship between coverage and words known (Hirsch & Nation, 1992). As coverage demands increase, vocabulary sizes become exponentially higher to meet them. This is why almost all technical word list studies have considered 95% to be a better threshold value (e.g. Coxhead, 2000; Coxhead & Hirsch, 2007; Konstantakis, 2007, 2010; Wang et al, 2008). It should also be noted that none of the studies which suggested a 98% lexical threshold (Hirsch & Nation, 1992; Hu and Nation, 2000; Schmitt, Jiang & Grabe, 2011) used technical corpora to calculate this figure. The most technical was the Schmitt, Jiang and Grabe study (2011) which used an article from an EFL textbook and from *The Economist*. The others used works of popular fiction.

It is possible to calculate how many words an L2 learner of English requires with these threshold figures in mind, if used in conjunction with corpus analysis. Nation (2001a) used the Carroll, Davies and Richman corpus (1971) to estimate that approximately 12,000 lemmas were required to obtain 95% lexical coverage. In later research, Nation (2006) revisited this word count following the construction of fourteen 1,000 word-family lists (frequency bands) from the British National Corpus (BNC) based on their relative frequency. Using these lists, he was able to test their reliability through comparison with other corpora and found a good level of rigour which suggested that these lists were representative of the English language as a whole. From this point, it was possible to calculate how many words an L2 learner of English would require to read a variety of different texts, including novels, graded readers and newspapers. Using a threshold of 98%, he concluded that approximately 8,000-9,000 word families were needed to understand such texts.

This estimate has been converged upon by similar studies. Laufer and Ravenhorst-Kalovski (2010) estimated 8,000 word families (with the inclusion of proper nouns) for the 98% lexical threshold to be obtained, with as few as 4,000-5,000 for the 95% threshold. However, their study was not corpus-based. It might seem that even a target vocabulary size of 8,000-9,000 words might present too difficult an impediment for an L2 learner of English to overcome (Milton, 2009), yet there are reasons why these studies might not properly represent such a student's needs. These corpora contain a number of different types of register, from formal through to informal writing in many different contexts. The L2 learner may not have need of a spectrum of registers dependent on their reasons for learning an L2. This is particularly important for ESP students.

The research considered thus far has shown that it is possible to estimate the number of words an L2 learner of English may require by using the idea of coverage, a lexical threshold and corpus data. However, these vocabulary size estimates have been made for general learners of English rather than those studying ESP. These L2 learners have need of a specialist lexicon for their subject specific studies. For these L2 learners, a more specific corpus analysis is required using literature from their discipline (e.g. Konstantakis, 2007) and general academic registers (Coxhead, 2000). Specialist L2 learners might still need access to non-technical vocabulary in the course of their language practice, but due to the limited amount of hours teachers have available to instruct their students (Milton, 2009), it becomes more expedient to concentrate on a specialist lexicon.

The GSL is a list of almost 2,000 words (section 2.6) which provides approximately 80% coverage of general English texts (Coxhead, 2000). Indeed, this 2,000 word benchmark is often cited as a necessary basis for gist understanding of the English language (Nation, 2001a), although a more recent paper asks that this be raised to the 3,000 word mark (Schmitt & Schmitt, 2012). However, the GSL provides only about 75% coverage of academic texts (Coxhead, 2000). With only 570 word families, the AWL provides a further 10% coverage on average in many different academic domains (Coxhead, 2011), making a total of 85% coverage when combined with the GSL. This is considerably lower than any acceptable lexical threshold, but obtained with roughly 2,570 word families. Most other technical word list studies (section 2.6) have attempted to bridge this gap between the amount of tokens accounted for by a combination of the GSL and AWL with the proposed lexical thresholds for comprehension. An exception to this is Ward (1999) who compiled an engineering corpus of approximately 1 million words from which he created an engineering word list. He concluded that only 2,000 words were necessary to obtain 95% coverage in specialised engineering texts for first year engineering students. He did not use the GSL or AWL as

he intended to increase the efficiency of vocabulary teaching by removing any words from the GSL and AWL which his corpus data showed to be absent from engineering texts. However, at only 1 million words, it may be argued that his sample size was insufficient to make such claims. Nonetheless, it was a demonstration of how specialist corpus studies can establish word lists with a reduced vocabulary burden for ESP students with a level of efficiency far better than learning the first 8-9k frequency bands (Nation, 2006). This also applies to building a supplementary specialist word list. Vocabulary size efficiency in achieving threshold coverage is critical.

2.4 Word technicality

It has already been discussed how words occur in the English language with differing frequency. Some, such as those found in the GSL, are amongst the most commonly used. Nation (2001a) differentiated between 4 different categories of words based on this frequency behaviour: high-frequency, academic, technical and low-frequency words. The purpose of defining words in such a way is for pedagogical reasons. There should be a cost/benefit analysis involved in the decision to teach items of vocabulary and there is a greater benefit in teaching high-frequency items (Nation, 2001b). Therefore, a teacher may concentrate on teaching the high-frequency words in English as they are most likely to be encountered by any student. Less time may be committed to teaching the low-frequency words as they are, by definition, a lot less common in language.

There has been a tendency to consider the most common 2,000 headwords in the English language as the high-frequency lexis (Nation, 2001a). This is mostly a result of the GSL being set at roughly this number, in addition to other research which helped shape this number as significant. One such study was conducted by Schonell, Middleton and Shaw (1956). They recorded and manually transcribed instances of spontaneous speech and some interviews of 2,800 semi-skilled Australian labourers. They used this data to build a corpus of approximately half a million words. Their findings suggested that only 209 word families were sufficient to provide 83.44% coverage of their corpus and that a language threshold of 95% could be obtained with only 1,600 word families. These numbers converge with the idea of there being 2,000 high-frequency words. However, more recent studies using the modern Cambridge and Nottingham Corpus of Discourse in English (CANCODE) corpus by Adolphs and Schmitt (2003) found that only 93.93% coverage could be obtained with these 1,600 word families. In fact, they required 3,000 word families to exceed the threshold coverage value of 95%. Cobb (2007) looked at 30 target words from each of the 1,000-3,000 frequency bands of the BNC. On comparison with a 517,000 extract from another corpus, he

found that words from each of these bands occurred with sufficient frequency to be considered as high-frequency. It is with evidence from research like this that Schmitt and Schmitt (2012) call for a re-categorisation of the high-frequency range to include the first 3,000 words in English.

Low-frequency vocabulary is easier to categorise. Using the Range programme (Nation & Heatley, 2002), Nation revisited his earlier work (Nation, 2001a) to change his recommendation for threshold coverage at 98% (section 2.3). Using a range of texts from different registers, he concluded that approximately 8,000-9,000 words are sufficient in English to obtain this threshold. His data showed a very strong drop-off point after this where each further 1,000 frequency band provided progressively less and less coverage. Using this data, Schmitt and Schmitt (2012) categorised low-frequency words as anything after this point.

Mid-frequency vocabulary is therefore those words which exist between the 3,000 high-frequency range and the 9,000 low-frequency range (Schmitt & Schmitt, 2012). It may be expected that academic and technical vocabulary may be found within this part of the spectrum. Using a Lextutor BNC-20 frequency analysis, they noted that 64.3% of the headwords from the AWL were within the 3,000 high-frequency range. However, the rest of the AWL did occur within the parameters of their mid-frequency vocabulary. Laufer and Ravenhorst-Kalovski (2010) found that university students in Israel required 6,000-8,000 word families to cover 98% of the examination reading texts in order to obtain a mark on a university entrance examination which showed they had sufficient comprehension to read academic material independently. Guided reading required knowledge of 4,000-5,000 word families, which allowed 95% coverage. Thus academic vocabulary may be found across the mid-frequency range and its knowledge is essential for L2 learners of English in a university setting. Receptive knowledge of this vocabulary will enhance a student's ability to understand academic writing and speaking, whilst a productive knowledge may help a student's ability to write and speak in such a register (Nation, 2008).

Academia is a large field with many widespread disciplines. Different subject areas may use the same words with different meanings and this leads to the issue of monosemic bias, as a homograph could misrepresent the composition of word families and hence affect the burden of learning such words (Wang & Nation, 2004). Whilst their study did not find this effect, one by Hyland and Tse (2007) did. The example given was that the word *process* was far more likely to be encountered as a noun by science and engineering students than by social scientists, who more frequently encounter it as a verb. This highlights the overlap between the categories of words which Nation (2001a)

considered academic and technical.

Technical words, which Nation (2001a) estimates as accounting for roughly 5% of words in a specialist text, are clearly subject dependent in meaning as well as frequency. Chung and Nation (2003) investigated the different means by which a word's technicality might be evaluated. They used an anatomy book of 5,500 tokens as their corpus data and used 4 different tools to identify the technicality of the vocabulary it contained: a rating scale, use of a technical dictionary, contextual clues and a corpus approach. Their 4 point rating scale was identified as the most accurate method of determining word technicality and hence used it as the means by which they would compare the relative success of the other approaches. Using this method they identified 227 terms which were considered of a technical nature. They found that when collocates were considered, the approach using computer frequency profiling obtained a 91.9% accuracy rate compared to their rating scale. When they used a technical dictionary to identify the same vocabulary they achieved a 73.9% accuracy rate. The clues-based method proved more accurate than using specialist dictionaries (83.1%), but was considered the least satisfactory as it accounted for only 135 of the 227 technical terms identified by their rating scale. This research also showed that technical vocabulary can be very frequent in a specialist text as it accounts for a significant proportion of the running tokens in the literature of a specialist field. They found a 31.2% coverage of technical lexis in their anatomy data compared to 20.6% in applied linguistics; far greater than Nation's estimate (2001a). With regards to their findings, the authors emphasised the importance of identifying technical vocabulary in domain-specific literature.

To summarise this section, it may be seen that words occur along a spectrum of frequency from high to low. High-frequency words may be considered to be those in the first 2,000-3,000 bands and low-frequency falling outside of the 8,000-9,000 bands. Academic and technical vocabulary may be found anywhere along this continuum, but mainly within the mid-frequency word range. If this vocabulary can be identified, then it can be of great practical use to an ESP student in their studies. The AWL is a list of the most common 570 academic headwords, but it only accounts for 10% of coverage in most academic texts (section 2.6). There remains the need for specialist lists to be built within each academic discipline to identify the subject specific technical vocabulary an ESP student must also learn. In addition to this, it is also interesting to note that corpus analysis provides one of the most accurate tools for identifying this lexis (Chung & Nation, 2003). If used in combination with a specialist dictionary then even greater levels of surety regarding the technicality of a word list may be obtained (Chung & Nation, 2003).

2.5 Multi-word units

One of the most significant findings from corpus research is that language consists not only of single words, but of multi-word units (Martinez & Schmitt, 2012). These multi-word units, or formulaic language as they are often known (Wray, 2002, 2008), are semantically opaque chunks of language which have a meaning separate and distinct from the sum of the individual words from which they are comprised. This is referred to as compositionality. There is gathering evidence that these multi-word units are acquired, processed, stored and produced by the L2 learner in a similar way to individual words and that they are a necessary part of language use (e.g. Schmitt, 2010; Wray, 2008). Martinez and Schmitt (2012), in their review of the evidence, highlight the most important research conclusions: phrasal expressions are ubiquitous; meanings and functions are often expressed by multi-word units; formulaic language has processing advantages; and they improve the impression of an L2 learners language capability.

Lexical chunks, such as *never mind*, *to stand up for* or *by all means* are so pervasive that they should be included in any EFL instruction (Martinez, 2013). It is important to note that some examples of multi-word units are idiomatic whilst others such as phrasal verbs are not. The intuitive understanding of this formulaic language by non-native speakers should not be overestimated. Ermann and Warren (2000) estimated that 58.6% of written and 52.3% of spoken discourse is made up of formulaic sequences. Biber (1999) placed the estimate at roughly half the value of the above authors, but this still accounts for a significant and important proportion of all discourse. Martinez and Murphy (2011) show that in a text where 95% coverage is obtained using the first 2,000 frequency bands from the BNC, if formulaic language were removed then coverage would drop to 61%. Martinez and Schmitt (2012) compiled a PHRASE list consisting of 505 multi-word items, which if integrated and calculated into the top 5,000 frequency bands, would account for 10% of all headwords. 97.8% of these multi-word items were in the first 2,000 frequency bands alone. This research in particular demonstrates the pervasive nature of multi-word units and their frequency within the high-frequency word category and it supports the idea that knowing individual words may not be sufficient to achieve adequate comprehension and hence can lead to misinterpretation and poor text comprehension (Martinez & Schmitt, 2012).

There is growing evidence which shows just how important multi-word units are in language use and therefore in second language acquisition. However, as an area of research they do present a

problem for any study which is attempting to compile a specialist word list. This is because many of the words which comprise multi-word units have been shown to stem from the high-frequency vocabulary range (Martinez & Schmitt, 2012). Despite the frequency of formulaic language in discourse and the high-frequency range of the individual words which multi-word units contain, neither the GSL or Nation and Heatley's (2002) BNC frequency bands contain any lexical chunks. However, much of the research on formulaic language is interested in general use as opposed to specialist use, but that does not mean that multi-word units with domain-specific meaning would not exist within the high frequency bands. The same applies to the AWL: it is also a single word list only, but may contain words which contribute to specialist lexical chunks. Any study designed to create a specialist word list which acts as a supplement to the GSL and AWL will encounter such issues, if it intends to include multi-word units.

In building a supplemental list to both the GSL and AWL it is common practice to eliminate any words from either of these lists from the specialist list. This is due to one of the criteria originally established by Coxhead (2000) in creating the AWL. She stipulated that items to be included should demonstrate specialised occurrence, meaning that they should not already be present in the GSL. Work which has built upon the AWL has further stipulated that AWL items were not to be included in their technical word lists (Konstantakis, 2007). However, eliminating words from the GSL and AWL introduces the risk of losing domain-specific multi-word units unless they are comprised of words from the technical list only. In Durrant's (2009) research into collocations in academic discourse, he differentiates between lexical and grammatical collocations. In his list of 1,000 lexical collocations, 425 of these items overlap the AWL. However, these collocations are not specialist usage outside of academic registers and could be used in the literature of disciplines as diverse as medicine and law. It is when these expressions are domain specific they become problematic for a study like this. Ward (2007) continued his research into ESP with a study of collocations in English in engineering literature using a 2 corpora totalling 630,000 tokens in size. He mainly identified multi-word units acting as noun phrases throughout this corpus. He concentrated on the wide range of collocations using the terms *gas*, *heat* and *liquid*. All these items hail from the GSL. Using vocabulary profiling software he was able to calculate the frequency with which these keywords collocated with others. For example, he found that the word *gas* showed collocational activity in 66% of its occurrences. Fraser (2009) studied collocational behaviour in a specialist pharmacology corpus of 500,000 tokens. He created a list of the 'Top 100' collocations he found and they included a mixture of what might be considered general, such as *et al* (Number 1), to the more specialised *heart failure* (Number 12) and *endothelial cell* (Number 14). Interestingly, only *heart failure*

includes items found in the GSL, but this does serve to demonstrate how the inclusion of multi-word units can be problematic when one is considering specialist lexis.

Multi-word units are clearly an important new area of research in SLA. New research emphasising corpus analysis techniques has highlighted the frequency with which they are found across all discourse. However, they do present methodological dilemmas for anyone wishing to incorporate them in a technical word list as they may consist of individual words which are subsumed by higher frequency lists. This is certainly an interesting area for further study in this field.

2.6 The GSL, AWL and specialist word lists

The General Service List (West, 1953) has come under criticism in many ways over the years since it was produced. These criticisms have questioned its age (Richards, 1974), its range (Engels, 1968) and its expandability (Gilner & Morales, 2008). It is interesting to note that the GSL is actually a reissue of the *Interim Report on Vocabulary Selection* (Faucett et al, 1936), and that the current version used (West, 1953) is no longer in print. It is worth mentioning that items picked for inclusion within the GSL were not picked on corpus frequency data alone. Some items were chosen by committee. Nonetheless, over a 100 years of corpus analysis data has shown that the highest frequency words account for the vast majority of all language used: the most common 2,000 words (such as in the GSL) have a coverage of between 70-95%, regardless of the source of the text (Gilner, 2011).

There seems to be disagreement regarding the number of words on the GSL. Nation and Hwang (1995) reported using a list of 2,147 word families, Nation (2004) reported 1,986 word families, and Gilner (2011) claims there are, "1,907 main entries and 3,751 orthographically different words (in principle, common derivatives and compounds)." However, further studies have compared the GSL with frequency bands from the BNC and a remarkable amount of similarity has been found. The GSL in combination with the AWL account for 88% of the 3 highest frequency bands from the BNC. Only 301 headwords were absent (Nation, 2004). Nation concluded that differences in distribution, coverage and content might qualify the need for a replacement to the GSL though it is uncertain as to how this might be done. Subsequent refinement of the BNC has indeed lessened these differences between BNC frequency data and the GSL (Gilner & Morales, 2008). It would seem that the GSL has withstood the rigours of age quite well and is still relevant to use in further studies. The majority of technical word list studies have continued to use it (Coxhead, 2000;

Coxhead & Hirsch, 2007, Konstantakis, 2007, 2010; Wang et al, 2008).

The Academic Word List (Coxhead, 2000) is a list of 570 word families, organised to level 6 of Bauer and Nation's scale (1993). The data was retrieved from a 3.5 million token corpus which was divided into 4 sections: arts, commerce, law and science. Each of these sub-corpora contained approximately 875,000 tokens. To build the corpus, Coxhead made use of 414 different texts which were balanced for length, but contained material from a number of different sources, such as textbooks, articles, book chapters and manuals. The intent was to produce a list of vocabulary which would be useful to any L2 English student in an English medium academic environment.

Four different selection criteria orchestrated the selection of material for the AWL. Firstly, as it was designed to be supplemental to the GSL, no words from the GSL were allowed to be in the AWL. Items were also chosen on the basis of their frequency, range and uniformity. Each word had to occur at least 100 times across the corpora (frequency) although this was later dropped to 80. It had to be present in at least 15 of the 28 subject areas (range) and over 10 times in each of the 4 sub-corpora (uniformity). The AWL was then divided into 10 sub-lists based on frequency (Coxhead, 2000).

According to Coxhead, the purpose of the AWL was to assist EAP teachers in a principled selection of vocabulary for their students to learn (Coxhead, 2011). It has been used by a number of different researchers in their ESP research and it has averaged at approximately 10% coverage across all this research (e.g. Cobb & Horst, 2004; Li & Qian, 2010; Martinez, Beck & Panza, 2009). However, it does not consistently do so as some academic disciplines are better represented than others. Coxhead, Stevens and Tinkle (2010) obtained only 7.05% coverage by the AWL in their investigation of an approximately 280,000 token corpus of secondary school science textbooks. Hyland and Tse (2007) obtained 10.6% coverage of their 3.3 million word academic corpus, but raised separate concerns regarding the way lexical items from the AWL occurred and behaved. They questioned the range, frequency, collocational activity and meaning of words in the AWL in a multi-disciplinary sense. There can also be issues of monosemic bias (section 2.4).

Nonetheless, the AWL remains relevant to modern vocabulary studies and this is why it is often utilised within the creation of specialist word lists (e.g. Coxhead & Hirsch 2007; Konstantakis, 2007, 2010; Wang et al, 2008). Coverage provided by the GSL and the AWL vary slightly in the research. Konstantakis (2007) covered 90.38% of all tokens in his research and Aichah (2012)

accounted for 91.58% between these 2 word lists. This still falls short of the lexical threshold value of 95-98%. For this reason, some researchers have chosen to build specialist word lists which act as a supplement to the GSL and AWL. Others, such as Ward (1999) with his Engineering Word List, have chosen to ignore them altogether.

Hirsch (2004) suggests that if a text has a low frequency of general words then it has a higher frequency of technical words. These are words outside the 3,000 BNC frequency band, but below the 8,000-9,000 bands (Schmitt & Schmitt, 2012). This is supported by Chung and Nation (2003), who reported the low frequency of general words and the high frequency of technical words in their study of anatomical and applied linguistics corpora. This phenomenon is the other reason which has led to researchers creating specialist word lists. As this study is interested in the requirement of a technical word list for the discipline of computer science, it is necessary to look in more depth at similar studies which have been conducted in other branches of academia.

Konstantakis (2007) created a Business Word List (BWL) which was designed to act in a supplemental fashion to the GSL and AWL. His research utilised the Published Material Corpus (PMC) by Nelson (2000). It is a 600,000 token corpus derived from 33 business English textbooks. His study led to the creation of a 480 word list, expanded to level 6 (Bauer & Nation, 1993), but only achieved a coverage of 92.93%, which fell short of his stated 95% coverage goal. In his unpublished PhD thesis, Konstantakis (2010) states that this might have been the result of the under-representation of both academic and business words in his original corpus. The PMC was not compiled from academic business English literature, but from business English textbooks. This would seem a fair assumption as the AWL only accounted for 4.66% coverage of his corpus, less than half of what it has consistently averaged elsewhere. Certainly, his second BWL (Konstantakis, 2010) which was based on a larger 1 million token corpus from a range of academic sources, achieved greater success. He obtained 95.62% coverage, yet his word list reached an incredible 1,613 word families. One further point of interest is that he confirmed the technicality of his list through its comparison with a second business corpus. This is a necessary research step to take, as a word list built from a corpus is likely to provide good coverage of said corpus (Coxhead, 2000). To ensure the representativeness of a word list it should be empirically tested against further literature from the field.

Coxhead and Hirsch (2007) compiled a science-specific word list. In this work, they criticised the poor representation of science vocabulary in the GSL. Using a 1.76 million token corpus they

extracted 318 word families which provided 4% extra coverage in science literature, but only 0.4% in non-science disciplines like arts, commerce, law and fiction. This was not a way to establish the technicality of a word used by Chung and Nation (2003). A technical word list should produce poor coverage of a general corpus as the language is less technical on average. However, whilst their study was designed to add to both the GSL and AWL, they did not report their findings on the total coverage obtained when all 3 lists are considered together.

Another example of a specialist word list, which utilises the GSL but not the AWL, is the Medical Academic Word List (MAWL) created by Wang, Liang and Ge (2008). They used a 1 million token medical corpus taken from 288 medical journal articles. Their MAWL contained 623 word families which gave an additional 12.23% coverage to the GSL which is a considerable amount of coverage for such a small word list. However, it is important to note that 55% of the MAWL contained lexical items found within the AWL, which only serves to highlight the academic robustness of the AWL. Again, the researchers failed to report the total coverage they obtained with the GSL and MAWL combined so it is unknown whether or not the lexical threshold for comprehension was established.

A final example of a technical word list is the Law Word List (LWL) created as part of a Master's dissertation by Aichah (2012). He compiled a Law Corpus (LC) of 3,843,107 tokens for his research. From this he obtained a total of 91.58% coverage using just the GSL and AWL and when he added his LWL he achieved 95.85% coverage which met Laufer's (1990) lexical threshold. There were 373 headwords in this list which accounted for the extra 4.27% of tokens covered. His study was noteworthy due to his attempt to create a technical multi-word list for law. He overcame the methodological problems (section 2.5) by allowing word families from both the GSL and AWL to be included in this list. His multi-word list did not contribute extra coverage for this reason, but at least demonstrated that corpus analysis can effectively identify the presence of such formulaic language (Chung & Nation, 2003).

This section has looked at the GSL, AWL and specialist word lists in some detail. Whilst there have been many criticisms aimed at the GSL over the years, it has proven surprisingly robust in this time and it shares a very high level of similarity with purely corpus-driven frequency lists (Gilner, 2011). This might say something about the fidelity of language over time, but most importantly it shows that it is still a valid piece of research upon which to build. Whilst the AWL has had its detractors, it also consistently performs at approximately 10% coverage across all disciplines studied to date (Coxhead, 2011). However, there are still academic disciplines against which the AWL has not been

tested, including computer science. As for other specialist word lists, most have chosen to make some use of either or both the GSL and AWL for these reasons. Those that have not, such as Ward's engineering word list (1999), did so for the sake of economy. It was found that an engineering word list alone could provide lexical threshold as it shared items from the GSL and AWL. Nonetheless, the GSL and AWL have survived some close scrutiny and continue to have an impact on vocabulary studies today. Their continued influence on technical word list studies is significant (e.g. Aichah, 2012; Coxhead & Hirsch 2007; Konstantakis, 2007, 2010; Wang et al, 2008).

2.7 Conclusion

This literature review has sought to highlight and explain areas of important, recent research into vocabulary and consider where further research is required. It has been shown how corpus analysis combined with an understanding of word frequency data has allowed for the possibility of a lexical threshold for comprehension to be postulated (Schmitt, Jiang & Grabe, 2011). Whilst precise figures are unobtainable, research has shown that anywhere between 95-98% coverage should maximise a reader's chances of sufficient text comprehension. The drawback with this approach is that the figure is somewhat arbitrary, non-absolute and subject to diminishing returns. However, it does serve to supply a goal for any specialist word list study, offering a more economical vocabulary requirement than the 8,000-9,000 words which might be required otherwise (Nation, 2006). It has also been shown how the technicality of a word may be reliably demonstrated: comparison of a specialist word list to other corpora made from literature within the same discipline; comparison of the same data to non-technical corpora; and comparison with a technical dictionary have all been employed in these studies (e.g. Aichah, 2012)

These methods of comparison are of particular note when considering multi-word units. Their ubiquity and prevalence in language has only recently been described (e.g. Durrant, 2009; Martinez, 2013; Martinez & Schmitt, 2012). Perhaps it is only because of the recent discovery of such a phenomenon that specialist word studies have, for the most part, largely ignored formulaic language, but any contemporary research in this area must redress such issues. However, as most word lists, including the GSL and AWL on which this study will build, do not factor for multi-word units, this creates a methodological problem which must be resolved. Nonetheless, the GSL and AWL both continue to dominate research in this field and have proven robust, reliable and consistent in most studies.

There has been no published research of vocabulary in the discipline of computer science. It is a viable subject area to study due to its influence, relative modernity and the global dominance of the IT industry. It is a noteworthy omission that a subject which has defined the millennium has received no treatment in the ESP sector thus far. There is a need for a CSWL as a pedagogical tool for L2 English-medium computer science students to achieve better comprehension within their studies.

3. Research Questions

Having looked at the literature in the previous chapter, a few areas for further research have been identified. Firstly, there has been no attempt made to create a specialist word list for the discipline of computer science to date. As this is a very large and important academic discipline, it is important that such a gap in the literature should be addressed. Currently, there is no knowledge of the distribution of word frequencies in this field, nor is there any knowledge of the technicality of the types of word used. Secondly, it is clear that there has been very little work done on multi-word units, particularly within specialist word lists. These omissions within the literature lend themselves to the development of the following research questions:

1. How much coverage do the GSL and AWL provide within a computer science corpus?
2. If the GSL and AWL do not provide 95% coverage of a computer science corpus, can a supplemental word list be created which helps the GSL and AWL exceed this threshold?
3. To obtain 95% coverage in a computer science corpus, how many word frequency bands are required? Is learning frequency bands more or less efficient than learning the GSL and AWL combined with a supplemental computer science word list?
4. Is the computer science word list a technical word list?
5. Are there subject specific multi-word units in computer science?

The following chapters of this paper will attempt to answer these questions as part of the research conducted for this study. As many previous studies have relied on corpus data to identify technical vocabulary within other subject areas (e.g. Aichah, 2012; Coxhead, 2000; Coxhead & Hirsch, 2007, Wang et al, 2008, Konstantakis, 2010), it was intended that this would be the research methodology of this paper. These corpora are digital collections of spoken or written discourse compiled with specific criteria to represent a body of language data for linguistic research (O'Keefe, McCarthy & Carter, 2007). Research by Verlinde and Selva (2001) compared corpus analysis as a means of gathering technical vocabulary data to that of expert intuition. They concluded that corpus analysis was the only means to gain empirical supporting evidence. As this was an empirical study, corpus analysis seemed the most appropriate methodological approach. The following chapter will thus outline the methodology involved in creating the corpora for study and in building the technical word list.

4. Methodology

4.1 Building the Computer Science Corpus (CSC)

The first process involved in answering the research questions (chapter 3) was to create the CSC. According to Sinclair (2005), there are several properties which a corpus should demonstrate and those most relevant to this study are size, balance and representativeness.

The size of the corpus represents the amount of linguistic data available for analysis. Similar studies to this paper were considered in section 2.6. In creating the MAWL, Wang et al (2008) used a corpus of 1,093,011 tokens. Konstantakis (2007) created a 600,000 token corpus for his technical business vocabulary study. A 1.76 million token corpus was used by Coxhead and Hirsch (2007) in their pilot study to create a scientific word list. However, a far larger study was originally conducted by Coxhead for the compilation of the AWL. She used a corpus of 3.5 million words to extract a broadly subject-independent list of academic words. It may be the case that a far larger corpus was required in order for claims about the frequency of general academic lexis to be made, but it is also true that a larger corpus allows for a better sampling of language. This follows Zipf's law (1935), which states that, "the frequency of any word is inversely proportional to its rank in the frequency table". A consequence of this is that, when frequency is a criteria for word selection in a technical word list (section 4.6), different size corpora require different frequencies of word occurrence. This is not necessarily a linear relationship as less frequent words do not appear in a smaller corpus. As this study was intended to create a supplemental word list to both the AWL and GSL, it was considered appropriate that a corpus of approximately the same size as that used by Coxhead (2000) would be sufficient and with well-defined frequency criteria.

There are also limiting factors for a study such as this, preventing the size of the corpus from being much larger. Although computers are powerful enough to handle far larger amounts of data, there were two further factors involved in restricting the size of the CSC, not pertaining to the current state of the technology needed to process it. Firstly, building a corpus requires a considerable amount of time and so the size of a corpus which can be produced for a study such as this is limited by the hours of labour available to build it. Secondly, to ensure balance and representativeness, the corpus-size was restricted by the availability of texts which could be properly assigned to each sub-discipline (explained later in this section). To this effect, it was decided that a corpus of 3.5-4

million words would be sufficient for the study.

Representativeness is the quality of a corpus through which it maintains textual relevance for the aims of a study. The CSC was designed to extract a list of words for use by students of computer science at university level. Therefore, it was necessary to select texts for inclusion which students in this discipline and at this level of study would use. Given the many sub-disciplines of computer science, it was important to ensure that each was properly represented within the corpus.

For the purposes of understanding what text types a computer science student might access, Coxhead and Hirsch (2007) interviewed university lecturers to obtain this information. The author of this paper also conducted similar interviews. Amongst the suggestions offered were textbooks, journal articles, special interest group newsletters and conference proceedings. Textbooks are not ideal for corpus building due to issues such as author bias. As textbooks are large texts, often with only one author, including them in a corpus can skew the results due to an author's preference for particular words and other idiosyncrasies (Atkins et al, 1992). This meant that the CSC was to be built using journal articles, special interest group newsletters and conference proceedings. These were shared across two corpora: journal articles and conference proceedings (which included the newsletters of special interest groups).

The question of identifying the sub-disciplines within computer science was a complicated issue. There is very little agreement within the discipline as to how it may be sub-classified. Considerable crossover exists between these sub-disciplines, such that any paper might overlap two or more different ones. A definitive list with papers organised by their primary sub-discipline was required and this was only possible through the use of the Association of Computing Machinery (ACM) and their digital library.

The ACM (www.acm.org) is the world's largest, not-for-profit dedicated computing organisation. It is a US based learned society for computing, which together with the IEEE Computing Society forms the world's largest academic resource in this discipline. They employ a Computing Classification System which identifies 11 major sub-disciplines of computer science. However, one of these sub-disciplines is the field of Applied Computing which concerns the use of computing within other academic subjects such as medicine and law. As these are not relevant to the current study and might cause technical lexis from other subjects to appear within the corpus with inflated frequency, this sub-discipline was omitted from the corpus. This meant that 10 sub-disciplines

remained viable for the study as seen in Table 4.1.

Computer science sub-disciplines	
Computer systems organisation	Mathematics of computing
Computing methodologies	Networks
Hardware	Security and privacy
Human-centred computing	Software and its engineering
Information systems	Theory of computation

Table 4.1 Computer science sub-disciplines defined by the ACM

Having looked at the corpus qualities of size and representativeness, it was necessary to consider balance. Maintaining this property ensures that a corpus has no bias to any of its sub-corpora by maintaining an equal distribution of equally sized texts throughout. This study had identified the 2 primary text types for inclusion within the corpus and the 10 different sub-disciplines it must contain. Therefore the CSC was to consist of 20 different sub-corpora. Due to the limiting factors of corpus size such as frequency variation and availability of sufficient texts per sub-discipline, a final size of the CSC was determined to be approximately 3.6 million tokens. This in turn meant that each of the 20 sub-corpora would contain about 180,000 tokens and that the CSC could be partitioned into 2 major corpora by text type of roughly equal size: the Computer Science Journal Article Corpus (CSJAC) and the Computer Science Conference Proceeding Corpus (CSCPC). Each of these would contain approximately 1.8 million tokens. The overall balance of the CSC can be seen in Table 4.2.

Sub-discipline	**Corpora**	
	CSJAC	**CSCPC**
Computer systems organisation	186662	180350
Computing methodologies	188002	179609
Hardware	185311	180748
Human-centred computing	184777	186486
Information systems	182366	179877
Mathematics of computing	180182	179532
Networks	185135	187321
Security and privacy	179608	181897
Software and its engineering	183266	186246
Theory of computation	183325	180637

Average	183863.4	182270.3
Total	**1838634**	**1822703**

Table 4.2 Number of tokens per sub-discipline in the CSJAC and CSCPC

4.2 Selecting texts for the CSC

The most important criterion for the selection of texts for the CSC was that the texts would all be digitised. In building a corpus of this size for the study it would have been unworkable to manually transcribe printed texts to a digitised version. The ACM was chosen to be the sole source of texts for the CSC as they have a comprehensive digital library. An electronic version of all texts for inclusion in the corpus was available for download from this library via the Athens cookie. Also, the ACM has developed the Computing Classification System (http://www.acm.org/about/class/1998), which they employ to attribute a primary, secondary and even tertiary level of nomenclature to each text based on the sub-disciplines it represents. This provided a simple system for the selection of each text considered for the CSC. A text could only be chosen to go within any of the sub-disciplines, if that was its primary classification. For example, a journal article might be indexed as having a primary classification of 'Networks' with a secondary classification of 'Hardware'. In this instance it would be included within the Networks corpus not the Hardware corpus. This helped ensure both the representativeness and balance of the CSC and avoid any accidental duplication of material. In a similar manner, all texts are clearly marked by their text type which further helped in the building of a corpus with unique entries. Using this methodology for the selection of texts, the CSC comprised of a total of 165 journal articles and 243 conference proceedings (Table 4.3).

Sub-discipline	**Number of texts**	
	CSJAC	**CSCPC**
Computer systems organisation	17	24
Computing methodologies	13	25
Hardware	19	24
Human-centred computing	18	29
Information systems	13	24
Mathematics of computing	19	24
Networks	19	27
Security and privacy	15	23
Software and its engineering	15	23

Theory of computation	17	20
Average	16.5	24.3
Total	**165**	**243**

Table 4.3 Number of texts per sub-corpora

The vast majority of these texts had multiple authorship. Between a total of 408 different texts used within the CSC, there were in excess of 1,000 different authors credited with their writing (Appendices C and D). This further assists in maintaining proper balance and representativeness within the corpus whilst avoiding the bias of author idiosyncrasy (Atkins et al, 1992). In addition, variation in the size of each text was kept to a minimum, where possible, to further improve the validity of the corpus. As can be seen from Table 4.3, a considerable amount of consistency in the number of texts per corpus was achieved. On average, the length of each text in the CSJAC was 11,143.24 tokens, whilst the average length of each text in the CSCPC was 7,500.84 tokens. This is because journal articles are slightly longer texts on average than conference proceedings and special interest group newsletters. It was not possible to obtain an equal amount of texts in each corpus whilst maintaining equal size of the corpora as every text differs in size. However, overall there were significantly lower differences from the mean size obtained in the compilation of this corpus than in previous similar studies (e.g. Aichah, 2012; Wang et al, 2008). Given the large size of the corpus, any small effects caused by differences in text size are thus mitigated.

4.3 Editing texts for the CSC

All texts selected for the CSC were available for download as PDF files from the ACM. However, each text underwent a series of editing steps before they were ready for inclusion in the CSC as TXT files. The software which was used for this study (section 4.4) was AntWordProfiler (Anthony, 2008) and AntConc (Anthony, 2002) and they are only capable of handling TXT files.

Each PDF file was copied into a text editor. After this, the list of references, appendices, page titles, authors' names, keywords and content pages, copyright information, publication names and tabular data were removed manually from each TXT file. This is because such information was not considered to represent linguistic data or was highly repetitive in nature. The decision to remove tabular data was made during the editing of the corpus as it was noted that tables were more likely to contain information in the form of numerical, mathematical or programming data. As they were clearly delineated from the body text, they did not represent a linguistic element. Further

information such as those provided in the references and appendices were considered excess to the understanding of each text and not necessarily something which students would encounter in their domain specific studies. Other deletions such as page titles and author names occurred with relatively high frequency and so would have affected the overall results. Items such as the article title, abstract data and footnotes were all kept as these were seen as necessary for understanding of the text.

The next stage of the editing process was to use regular expressions to find and replace further aspects from the texts, as well as to prepare them for use with the software. The ACM uses a convention of placing in-text citations within square brackets like this []. As proper nouns do not contribute to the learning burden (Konstantakis, 2010), they may be edited from a corpus. Hence the regular expression \[.*?\] was employed to remove all in-text citations. It is important to note that whilst this removed all citations of the style [Name, Year], it only removed the date from citations such as Name [Year]. As a result, many names remained within the corpus and had to be dealt with separately (section 4.5). However, given that the ACM employ a strict system for referencing which students may easily follow, this part of the editing process was considered acceptable as it is separated from the linguistic data.

Once citations had been removed, formatting was stripped from the texts and line-breaks removed. It was important to remove all non-alphabetic data from the TXT files as the software is not capable of handling this information. This was achieved using the regular expression [^a-zA-Z]. It replaces all characters not in the Roman alphabet. Hence all numbers, punctuation and other special characters were deleted. Finally, any extra white space caused by the excision of non-alphabetic characters was removed. Once this step had been completed, the TXT file was added to the appropriate sub-corpus.

After all texts had been edited and saved in this fashion, the corpus was complete. It totalled 3,661,337 tokens in length, divided evenly between the 2 major text types and between all 10 sub-disciplines of computer science. In this way a representative, balanced corpus of sufficient size for identification of technical lexis in the field of computer science was obtained (Table 4.4).

Sub-discipline	Corpus length
Computer systems organisation	367012
Computing methodologies	367611

Hardware	366059
Human-centred computing	371263
Information systems	362243
Mathematics of computing	359714
Networks	372456
Security and privacy	361505
Software and its engineering	369512
Theory of computation	363962
Average length	366133.7
Total length	**3661337**

Table 4.4 Corpus length per sub-discipline in tokens

4.4 Software

The software which was selected for analysis of the corpus data was AntWordProfiler (Anthony, 2008). This is available for free download from Laurence Anthony's homepage (http://www.antlab.sci.waseda.ac.jp/index.html). It is a powerful corpus analysis tool capable of extracting frequency and range data from a number of large corpora simultaneously. It accomplishes this through comparison of the corpora with a variety of word lists. It is pre-loaded with the GSL1, GSL2, and AWL lists. However, it is possible to replace or add to these lists with those from other sources.

It was decided from the outset that a different set of word lists would be used. Paul Nation's homepage (http://www.victoria.ac.nz/lals/about/staff/paul-nation) includes the Range programme (Nation & Heatley, 2002) for free download with a number of different word lists available including the GSL and AWL, the first 14k frequency bands of the BNC and the first 25k frequency bands of the BNC/COCA (Corpus of Contemporary American English). A comparison of the GSL and AWL lists from both sources was made by checking coverage of one set of lists with the other. Nation's lists had a greater level of expansion and also included both British and American spellings. This was important as the corpus literature was all acquired from a US source, yet this research is intended to meet the requirements of UK-based computer science students. Therefore corpus analysis for this study was to be performed using the AntWordProfiler tool (Anthony, 2008) with the Nation GSL and AWL word lists (Nation & Heatley, 2002).

This software can display vocabulary data by frequency or range. It can organise by token, type or group (word family). Importantly it can display this data by word list and it is also possible to view tokens which occur outside of the defined word lists. This component was essential in the process of extracting technical language for use within the CSWL (section 4.7).

AntConc (Anthony, 2002) is another useful programme for the analysis of corpus data. Of most use in the research conducted for this study was the *Collocates* tool. This functionality is capable of establishing the range and frequency with which tokens collocate to a number of positions left and right of the search term. It allows for the detection of multi-word units within a corpus and thus was a necessary tool for the final research question of this study (chapter 3).

4.5 Clearing unwanted data from the CSC

Before it was possible to extract the technical language from the completed CSC, it was necessary to add any unwanted data to separate word lists. This data included abbreviations, acronyms and proper nouns such as names. In order to identify the unwanted data, the CSC was run against the complete BNC/COCA 25k frequency bands (Nation & Heatley, 2002). Any tokens which fell outside this large frequency range were likely to include this unwanted data. Included within this profile were 2 word lists called Basewrd31 and Basewrd34 which were lists compiled for the purpose of collecting such information.

Basewrd34 from Nation's BNC/COCA 25k frequency bands (Nation & Heatley, 2002) contains a large list of common abbreviations and acronyms (such as *BBC*) which are often excluded from corpus token counts (Coxhead, 2000) as it is assumed that these are self-explanatory and place no learning burden on the reader (Konstantakis, 2010). In academic writing most unusual abbreviations are fully expanded into their constituent words the first time they are printed. It was into this word list that further acronyms or abbreviations were placed, that were found in the CSC but were not already in Basewrd34 (such as *RAM* or *JPEG*). Any other tokens which were not recognisable as English or proper nouns were also added to the list (such as *BitVec* and *fx*). These predominantly came from mathematics, algorithms and programming languages present in the corpus. The number of types and the coverage provided by Basewrd34 of the CSC before and after it was cleared may be seen in Table 4.5.

Word List	Before clearance of the CSC		After clearance of the CSC	
	Types	**Coverage**	**Types**	**Coverage**
Basewrd31	22409	0.99%	22448	1.07%
Basewrd34	1149	0.81%	1524	2.17%
Total	**23558**	**1.80%**	**23962**	**3.24%**

Table 4.5 Basewrd31 and Basewrd34

The BNC/COCA 25k frequency bands (Nation & Heatley, 2002) also contain a word list called Basewrd31 which is an extensive collection of proper nouns. Place names, company names and people's names are included within this text. Once the CSC had been cleared of proper nouns which were not already in Basewrd31 (such as *Google*, *Microsoft* and *Twitter*), the cumulative effects of its expansion could then be calculated (Table 4.5).

Overall, 3.24% of the corpus contained this kind of data. As with other studies of this kind (Coxhead, 2000), the decision was made to exclude all such data from coverage counts so that only the linguistic data which could cause comprehension problems remained within the corpus for further evaluation.

4.6 Criteria for the selection of technical words in the CSWL

Once the CSC had been cleared of unwanted data, it was then possible to start the compilation of the CSWL which was central to the last four research questions posed in chapter 3. Three criteria were enforced for the selection of technical words to add to the CSWL. These were specialised occurrence, range and frequency and were used by Coxhead (2000) in the creation of the AWL. Other researchers have also applied and modified these criteria in their research (e.g Coxhead & Hirsch, 2007; Konstantakis, 2007; Wang et al, 2008).

Specialised occurrence refers to the property of not existing within the GSL as defined by Coxhead (2000). Further research by those who built upon the AWL expanded this definition to include words which did not exist within the AWL (Konstantakis, 2010). This was considered a necessary criterion for this study as the CSWL was intended to act supplementally to both the GSL and AWL in the same way as the BWL (Konstantakis, 2007, 2010). AntWordProfiler (Anthony, 2008) ensured that this criterion was trivial as it can automatically filter words from the AWL if loaded with the appropriate word list. However, it was discovered during this process that some of the expanded

forms of words in both the GSL and AWL present in the CSC had not already been added to these lists. As a result, some further expansion of the lists was required whilst ensuring that they were only expanded to level 6 of Bauer and Nation's word families (1993). Some examples of this may be seen in Table 4.6.

GSL1	GSL2	AWL
Attack	Compete	Compute
Attacker	Competitive	Computes
Attackers	Competitively	Computerise
Map	Parallel	Computerize
Mapped	Paralleling	
Mapping	Paralleled	

Table 4.6 Examples of expanded forms added to the GSL and AWL

The next criteria which Coxhead established for selecting technical lexis was range (Coxhead, 2000). For a word to be included within a technical word list it must be present in at least half of the sub-corpora (Wang et al, 2008). This was a slight modification of Coxhead's (2000) original range, but one used by most studies since (e.g. Aichah, 2012; Coxhead & Hirsch, 2007). At this point the corpora for each text type had been combined so that there were 10 sub-corpora in use within the CSC: one for each sub-discipline of computer science. This meant that a word had to appear in at least 5 of these 10 sub-corpora for it to be included within the CSWL. Range is a useful property as it helps to ensure the representativeness of a technical word throughout the corpora.

Finally, there was the frequency criterion. This relates to the number of individual occurrences of a word in the corpus. For the AWL study, Coxhead (2000) initially set the minimum frequency at 100. However, this was lowered to 80 by the completion of her research as it was found to give a significantly better coverage. As the CSC was of a similar size to the one which Coxhead employed to create the AWL (Coxhead, 2000), it was decided that the same minimum frequency of 80 would be adopted. Other studies have scaled down this minimum frequency level to suit the size of the corpora they used for their study. In the MAWL (Wang et al, 2008) the minimum frequency was set at 30 for a 1.09 million word corpus and Coxhead and Hirsch (2007) used a minimum frequency of 50 for their 1.76 million word corpus. These relative frequencies have been calculated on the assumption of a linear relationship between frequency and corpus size, but this is only applies to Zipf's law (1935) in an infinitely long corpus. At smaller corpus sizes, words leave the frequency

tables altogether implying a non-linear relationship. For this reason the CSC was chosen to be approximately the same size as the one used by Coxhead (2000) so that the same frequency criterion could be used.

4.7 Completing the CSWL

Using fully expanded word lists for GSL1, GSL2 and the AWL, AntWordProfiler (Anthony, 2008) was again employed. It was possible at this stage to identify all words outside of these 3 word lists, ordered by range then frequency. Any word with a range of 5 and frequency of 80 or more was then added to the CSWL.

This list was then expanded to Bauer and Nation's (1993) word family level 6 (Table 2.1). As with words in the GSL and AWL in the above section, this was a necessary process for a number of reasons. Firstly, all word lists used by the software are expanded in such a way, so it was sensible to maintain this agreement. Any deviation from this standard might have influenced the statistical data. Level 6 expansion has been used as a basis by most of the research into technical word lists and so it would not be sensible to do otherwise (e.g. Coxhead, 2000; Konstantakis, 2007). Finally, level 6 expansion is considered to be within the morphological capability of students who must have achieved a certain level of English proficiency to be allowed to study computer science at a British university. The process of expanding these words was simplified by retrieving their individual entries from the BNC/COCA 25k frequency bands produced by Nation (Nation & Heatley, 2002).

One further stage of filtering was required at this stage. During the editing process of preparing TXT files for the CSC (section 4.3), all non-alphabetic characters had been replaced with a single white space. This included hyphens, so all previously hyphenated words in the texts were considered as separate tokens in the corpus. However, this had been anticipated and was one of the reasons the AntConc (Anthony, 2002) software had been selected. Any words which met the selection criteria for inclusion in the CSWL could be checked using the *Collocates* tool of this software to show whether it existed as a separate token or if it occurred only as part of a multi-word unit (including hyphenated words). If it occurred as part of a multi-word unit, it would qualify for the Computer Science Multi-Word List (CSMWL) if range and frequency demands were satisfied. In this way, many types which act only as affixes were removed from the CSWL (Table 4.7).

Affixes excluded from the CSWL	
non	*micro*
multi	*mega*
re	*poly*
co	*kilo*
meta	*giga*

Table 4.7 Affixes excluded from the CSWL

Following the successful completion of this process the CSWL was considered ready for testing. It consisted of a total of 433 word families, which were expanded into 1,943 types. A list of the headwords of the completed CSWL may be found in Appendix A.

4.8 Conclusion

The methodology outlined in this section covers all processes involved in the creation of both the CSC and CSWL. It was informed, at all stages, by previous research into technical word lists (e.g. Aichah, 2012; Coxhead, 2000; Coxhead & Hirsch, 2007; Konstantakis, 2007, 2010: Wang et al, 2008). It is believed that the CSC is an appropriately sized corpus which is both balanced and representative of the literature which a computer science student in the UK might expect to encounter during their studies. Furthermore, it was constructed in a systematic and transparent fashion to ensure it contained only unique entries from literature gathered from all the sub-disciplines within the field of computer science.

With the CSWL complete it was possible to consider the results of this research and answer the research questions posed in chapter 3. The results of this research are documented in the next chapter of this paper.

5. Results

5.1 Coverage of the GSL and AWL in the CSC

Previous research has diverged in its treatment of proper nouns and other unwanted data from its findings. Whilst Coxhead (2000) decided to exclude all such data from her results with the AWL, her later research with Hirsch (2007) and the work by Konstantakis (2010) included them. Wang et al (2008) made no mention of their treatment of this issue. In order for this study to be fairly compared to all these studies, it was necessary to report the findings with proper nouns both included and excluded. Also, given the nature of computer science literature with its significant proportion of mathematical, algorithmic and programming data, it was also decided to report coverage with all single alphabetic data (the letters except for *A* and *I*) removed from the GSL1 word list (Table 5.1). AntWordProfiler (Anthony, 2008) was used to obtain all these results.

Section 4.5 showed how 3.24% of the corpus data comprised of proper nouns and other unwanted data. Calculating the coverage of the GSL and AWL with this data excluded involved the following equation:

Coverage after exclusion = Coverage before exclusion / (100 - Unwanted Data) x 100

Word List	Coverage		
	Proper nouns included	**Proper nouns included and single letters excluded from GSL1**	**Proper nouns excluded**
GSL1	68.44%	64.34%	70.73%
GSL2	5.52%	5.52%	5.70%
AWL	12.38%	12.38%	12.79%
Total	**86.34%**	**82.24%**	**89.22%**

Table 5.1 Coverage of the GSL and AWL in the CSC

These results answer the first research question, posed in chapter 3. It may be seen that, if proper nouns data are included in the count then the GSL and AWL together provide only 86.34% coverage. This would mean that the CSWL would have to provide between 8.66-11.66% extra coverage to match the lexical threshold for text comprehension of 95-98% (Laufer, 1989; Nation, 2008). However, if proper nouns were to be excluded from the count (section 4.5), then the

coverage required by the CSWL to reach the threshold would need to be a lower 5.78-8.78%. Results from other studies show the difficulty of this objective (Table 5.2). The combined total of the GSL and AWL accounted for 91.58% of all tokens in the LC which Aichah (2012) compiled for his LWL study.

Word List	CSC	LC (Aichah, 2010)	PMC (Konstantakis, 2007)
GSL1	70.73%	77.71%	80.26%
GSL2	5.70%	4.08%	5.46%
AWL	12.79%	9.79%	4.66%
Total	**89.22%**	**91.58%**	**90.38%**

Table 5.2 Coverage of the GSL and AWL in the CSC, LC and PMC

5.2 Coverage of the CSWL in the CSC

Given the relatively poor coverage of the combined GSL and AWL within the CSC, it was clear that the CSWL would have to provide a significant amount of coverage to obtain lexical threshold values (Laufer, 1989). This data was collected in order to provide an answer to question 2 of this study (chapter 3). AntWordProfiler (Anthony, 2008) was again the software used to extract these results (Table 5.3).

Word List	Proper nouns included in coverage	Proper nouns excluded from coverage
GSL1	68.44%	70.73%
GSL2	5.52%	5.70%
AWL	12.38%	12.79%
CSWL	5.81%	6.00%
Proper Nouns (Basewrd31)	1.07%	
Abbreviations (Basewrd34)	2.17%	
Total	**95.25%**	**95.11%**

Table 5.3 Coverage of the CSWL in the CSC

So it may be seen that it was possible to obtain the lexical threshold for comprehending a text as stipulated by Laufer (1989). However, it was not possible to achieve the higher threshold of 98% suggested by Nation (2008). Given that a total of 3.24% of the corpus data consisted of proper

nouns and abbreviations, this threshold of 98% was not an achievable figure if unwanted data was maintained within the count. Hence, all further coverage data will be expressed with proper nouns excluded. As each corpus has a different proportion of proper nouns and unwanted data, the calculations for each had to be made separately.

The results thus far have shown a coverage value for the entire CSC only. In order for the representativeness of the CSWL to be ascertained, it was important to determine what coverage it provided by text type (Table 5.4). The 2 text type corpora which were used for this analysis were the CSJAC and CSCPC as described in section 4.2.

Word List	Coverage in the CSJAC	Coverage in the CSCPC
GSL1	70.40%	70.33%
GSL2	5.60%	5.74%
AWL	13.32%	12.13%
CSWL	6.02%	5.93%
Total	**95.34%**	**94.13%**

Table 5.4 Coverage of the CSWL in the CSJAC and CSCPC

Table 5.4 shows that whilst it was possible to maintain over 95% coverage with the CSWL in the CSJAC, this was not quite possible with the CSCPC. The main reason for this would appear to be the drop in coverage provided by the AWL, rather than a significant decrease in performance by the CSWL. The AWL provided 1.19% less coverage of conference proceedings than it did of journal articles whereas the CSWL provided 0.09% less.

Having looked at the coverage of the CSWL against the CSC, CSJAC and CSCPC it is also necessary to show the results obtained for coverage of the CSWL against the 10 sub-disciplines of computer science as defined by the ACM and used in this study (section 4.1). The results may be seen in Table 5.5.

Sub-Discipline	GSL1	GSL2	AWL	CSWL	Total
Computer systems organisation	67.96%	5.67%	14.20%	7.35%	**95.18%**
Computing Methodologies	70.43%	5.50%	13.52%	5.74%	**95.19%**
Hardware	68.22%	6.53%	12.98%	6.68%	*94.41%*
Human-centred computing	71.93%	5.49%	12.86%	4.18%	*94.46%*

Information systems	70.15%	5.70%	13.91%	5.84%	**95.60%**
Mathematics of computing	73.71%	5.41%	10.20%	6.15%	**95.47%**
Networks	69.74%	5.69%	12.76%	7.31%	**95.50%**
Security and privacy	70.16%	5.65%	13.17%	6.14%	**95.12%**
Software and its engineering	70.20%	6.20%	13.55%	5.67%	**95.62%**
Theory of computation	74.93%	5.15%	10.71%	5.08%	**95.87%**

Table 5.5 Coverage of the CSWL by sub-discipline

Out of the 10 sub-disciplines of computer science, it was possible to obtain the lexical threshold of 95% (Laufer, 1989) with 8 of them. It was not possible to achieve 95% coverage within the Hardware sub-discipline. This would appear to be due to the low coverage provided by GSL1 and the AWL in this instance. However, the same can not be said about Human-centred computing. There was a decrease in the coverage provided by the CSWL for this sub-discipline which would account for the drop below 95%.

Overall, the results found in this section tentatively support the finding that it is possible to obtain the minimum lexical threshold for comprehension of computer science literature using a combination of the GSL, AWL and CSWL. This addresses the second research question posed in chapter 3. It is important to note that this only applies to the CSC as this stage of the research. The claim that the GSL, AWL and CSWL can provide 95% coverage of further literature in the field of computer science can not be substantiated until they are compared to a second corpus of computer science texts (section 5.4)

5.3 Coverage of the CSC with the BNC and BNC/COCA word lists

The third research question of this paper (chapter 3) asked whether or not it would be more efficient to learn the GSL, AWL and CSWL combined, than to learn all the words from the frequency bands required to obtain 95% coverage of a computer science corpus. This requires the simple process of counting the combined number of word families within the GSL, AWL and CSWL and comparing that number to the number of frequency bands required to obtain the same coverage. If the GSL, AWL and CSWL provide a lower word family count, it may be said to be more efficient.

As mentioned in section 2.6, there is variation in the number of word families attributed to the GSL. For the purposes of this research, the number proposed is that taken from the GSL1 and GSL2 word

lists used (Nation & Heatley, 2002). GSL1 contains 1,001 word families, whilst GSL2 contains 988. The AWL has a total of 570 word families and there are 433 in the CSWL giving a total of 2,992 word families. In order for it to be equally efficient to learn vocabulary through the frequency bands, 95% coverage of the CSC would need to be obtained by the first 3k frequency bands.

Nation's Range programme (Nation & Heatley, 2002) comes with a number of different frequency bands available for download. One example gives the first 14k frequency bands extracted from the BNC only. The other one of note contains the first 25k frequency bands taken from the BNC/COCA corpora. When the first BNC 14k frequency bands were compared, a coverage of 95.65% was obtained by the first 10k frequency bands. The first 9k frequency bands only provided 94.94% coverage in total. Only 96.44% coverage could be achieved with all 14k frequency bands. Using the BNC/COCA word lists, it was only possible to acquire a total of 93.53% coverage using all 25k frequency bands. The discrepancy between these figures is possibly explained by the difference between the corpora from which the word lists were originally extracted. It also indicates the high proportion of low-frequency words used in the computer science discipline. Nonetheless, it may be seen that the first 10k frequency bands of the BNC are required to obtain the lexical threshold of 95%. This is considerably less efficient than learning the word families from the combined GSL, AWL and CSWL in answer to research question 3 (chapter 3).

5.4 Technicality of the CSWL

Schmitt and Schmitt (2012) defined mid-frequency vocabulary as that which falls outside of the first 3k frequency bands, but before the 8-9k frequency band. This is where they expected academic and technical lexis to be found. Using the 14k COCA frequency bands (Nation & Heatley, 2002), it was possible to check the distribution of words in the GSL, AWL and CSWL. The results may be seen in Figure 5.1. It is evident from this graph that the majority of words in the CSWL are to be found in the mid-frequency range, supporting Schmitt and Schmitt's theory (2012).

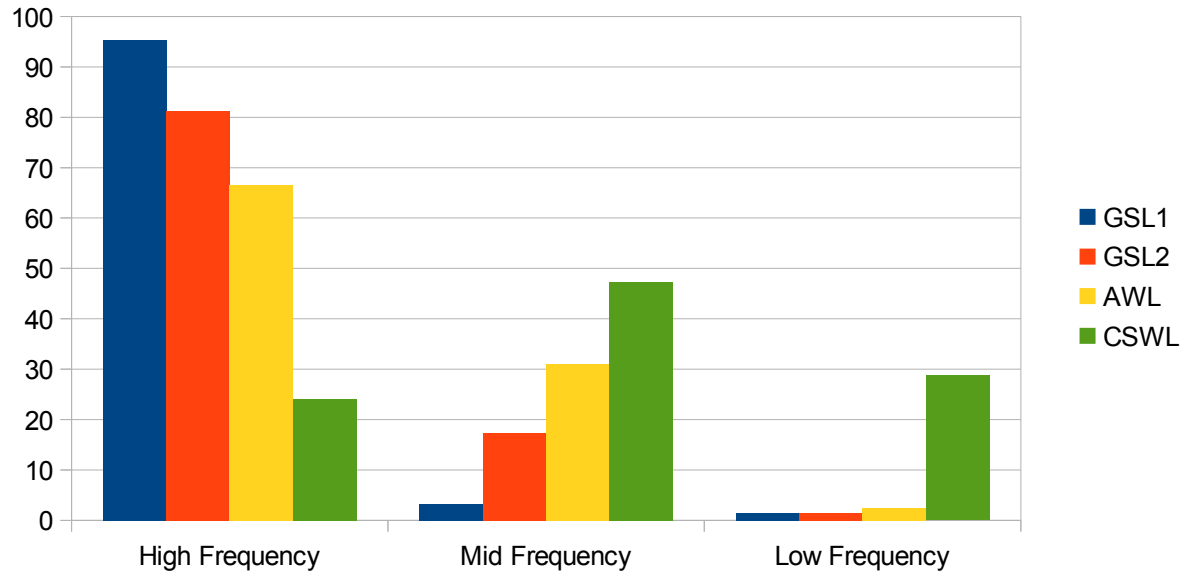

Figure 5.1 Distribution of the GSL, AWL and CSWL across the BNC 14k frequency bands

This was not sufficient evidence to claim the CSWL to be a technical word list. In order for this claim to be supported, there were 3 further research steps required. Firstly, the CSWL needed to be tested against a second computer science corpus. Coxhead (2000) noted that any word list derived from a corpus would be expected to perform well within that corpus. To establish the viability of the CSWL, further subject specific corpora-based testing was essential. Secondly, the CSWL needed to be tested against a fiction corpus. This would show the technicality of its lexis or if not, demonstrate a more general vocabulary with lower frequency items. Finally, Chung and Nation (2003) recommended the use of a specialist dictionary to check the technicality of a word list. The following sub-sections will discuss the methodology involved and results obtained.

5.4.1 Comparison against another computer science corpus

In order to test the technicality of the CSWL, it needed to be tested against a further computer science corpus. This second corpus was compiled using a different source of literature, combining the previous text types used for the CSC with textbooks. The reason for the inclusion of textbooks was that they are frequently encountered by students in the course of their studies and were originally recommended as a source of literature for students (section 4.1). Whilst they may skew the results of any corpus intended to build a representative word list (Atkins et al, 1992), it was envisaged that they would provide similar coverage data. The corpus was edited with the same

methodology as described in section 4.3. A corpus of 693,551 tokens was compiled from 23 different texts (Appendix E) and included a total of 16,787 lexical types. This was comparable in size to the test academic corpus built by Coxhead (2000) in the compilation of the AWL, which contained approximately 678,000 tokens. Interestingly, there were 32,539 different lexical types in the AWL test corpus which is over double the number of types in the computer science test corpus despite being of a slightly smaller size.

The CSWL's coverage of the smaller test corpus was 4.68% (Table 5.6) which was 1.32% less than the coverage it obtained within the CSC. Only 409 of the 433 word families of the CSWL were present in this test corpus so it might be said that the CSWL is not thoroughly representative of subject specific technical lexis. However, as a supplementary word list, it does provide a larger amount of coverage per word family than other similar studies. Konstantakis (2007) only achieved 2.55% coverage with 480 word families. It is also important to note that profiling some of the compound headwords from the CSWL, resulted in a greater frequency of multi-word or hyphenated representations than single token versions of a headword. For example, *dataset* only appears 30 times in the test corpus but 77 times as *data set.* Accumulated totals of these variations in spelling convention might push coverage in excess of 95% for this test corpus. The AWL continued to show a high level of coverage (12.28%) as it has done throughout this study.

Word List	Coverage of the CSC	Coverage of the test computer science corpus
GSL1	70.73%	72.28%
GSL2	5.70%	5.17%
AWL	12.79%	12.28%
CSWL	6.00%	4.68%
Total	**95.11%**	**94.41%**

Table 5.6 Coverage in the CSC and test computer science corpus

Despite not achieving threshold coverage (Laufer, 1989), these results do demonstrate a level of technicality in the CSWL. This provides further evidence that the CSWL is a technical word list as per question 4 of this study (chapter 3).

5.4.2 Comparison against a fiction corpus

To further investigate that the CSWL was a technical word list, not a general word list of lower frequency items, it was also necessary to test it against a corpus from a different genre. A corpus of fictional and philosophical literature was compiled for this part of the research. The resources were obtained from Project Gutenburg (http://www.gutenberg.org/), which is an online repository of books which have exceeded their copyright and are thus freely available for download. The intention of this part of the research was to demonstrate the frequency of occurrence of word families from the CSWL in a fiction corpus. As with Coxhead (2000), the number of texts, the length of the texts and the range of vocabulary items across the corpus was unimportant.

It may be argued that literature taken from this resource is unlikely to contain many items within the CSWL. This literature is at least 50 years old and pre-dates the rise of the computer industry. This might be a fair assertion as the CSWL contains some modern nouns (e.g. email and smartphone) which did not exist when these books were written. However, it contains a majority of words which cannot be dated in such a fashion and so it remained necessary to test the CSWL in this way.

The fiction corpus contained 3,671,673 tokens collected from 26 different texts (see Appendix F for references). The CSWL accounts for approximately 0.39% of all tokens within this fiction corpus, which is much lower than the 6% it covered in the CSC. Only 303 of the 433 word families from the CSWL were present (Table 5.7).

Frequency of occurrence	**Number of CSWL word families**	**Number of AWL word families (Coxhead, 2000)**
Not in fiction corpus	133 (30.72%)	30 (5.26%)
In corresponding technical corpus:		
4 times or more than the fiction corpus	240 (55.43%)	380 (66.66%)
3 times or more than the fiction corpus	15 (3.46%)	34 (5.96%)
Twice as often or more than in the fiction corpus	19 (4.39%)	52 (9.12%)
Less than twice as often in the fiction corpus	12 (2.77%)	52 (9.12%)
Less than in the fiction corpus	14 (3.23%)	22 (3.86%)

Total	**433**	**570**

Table 5.7 Occurrence of the CSWL and AWL in their appropriate technical and fiction corpora

Of the CSWL families, 373 are technical according to Coxhead's (2000) definition. That is to say, they are either 4 times as frequent in the CSC than in the fiction corpus, or not in the fiction corpus at all. This accounts for 86.15% of all word families in the CSC. Only 71.92% of Coxhead's (2000) word families were located in this group. A further 34 word families of the CSWL were at least twice as frequent in computer science texts as in fiction. The remaining 26 word families might be considered more general vocabulary. These results would seem to support the technicality of the CSWL in answer to research question 4 (chapter 3).

5.4.3 Comparison against a technical dictionary

The final area of research into the technicality of the CSWL involved comparing it to a technical dictionary. Just as the process of checking the CSWL against a second computer science corpus partially identified the subject specific nature of its lexis, Chung and Nation (2003) also recommended comparison to a technical dictionary as a valid means of supporting the same data. Konstantakis (2010) used this method with his BWL to demonstrate that more than half of the headwords on his list appeared in a business dictionary.

For the current study, the Oxford Dictionary of Computing (ODOC) (2008) was chosen. This dictionary has over 6,500 entries so is sufficiently large for this type of research. Chung and Nation (2003) argued that a larger dictionary is not necessarily better as it would contain a larger proportion of non-technical lexis. Any lexical type from the CSWL which had a main entry in the ODOC was noted. If a type from the CSWL appeared as part of a main entry in the ODOC, this was also recorded (Table 5.8).

	Entry in subject specific dictionary		
	Main entry	**Part of a multi-word unit**	**No entry**
CSWL	229 (52.9%)	75 (17.3%)	129 (29.8%)
LWL (Aichah, 2012)	139 (37.3%)	77 (20.9%)	156 (41.8%)
BWL (Konstantakis, 2010)	686 (54%)	-	580 (46%)

Table 5.8 Distribution of words in a technical dictionary from this study and related research

A total of 70.2% of all word families in the CSWL had an entry in the ODOC. This lends some support to the notion of it containing mostly subject specific technical lexis. For his LWL, Aichah (2012) identified only 58.2% of words with dictionary entries for his subject whilst Konstantakis (2010) only 54%. Although the latest edition of the ODOC was used for this study, some of the items on the CSWL do not appear as they represent more modern terminology (e.g. smartphone). This is a potential cause for concern for the durability of this research as this demonstrates that the vocabulary of the CSWL may be more transient. However, these results do support the subject specific technicality of the CSWL in answer to the 4th research question (section3).

5.5 Multi-Word Units in the CSC

The final research question of this study asked if there were any subject specific multi-word units in computer science literature (chapter 3). Recent research has demonstrated the ubiquity of these expressions in language (Martinez & Schmitt, 2012). In order to ensure these phrases were semantically transparent with a meaning separate to the sum of their individual words, it was decided that any multi-word unit found within the CSC would have to have an entry within the ODOC. This also ensured that such expressions were lexical rather than grammatical in nature (Durrant, 2009).

The software used for this part of the research was AntConc (Anthony, 2002). It had many useful tools to assist in this process. The *ngram* tool could identify any chunks of tokens that existed within a corpus and the frequency with which they occurred. Once these chunks had been identified they could then be processed by the *Concordance Plot* tool which noted the corpora in which it occurred, thus identifying the range of such chunks. This was considered necessary as any multi-word unit, which was to be added to the CSMWL, should also meet the same criteria that were used in building the CSWL (section 4.6). They should have a frequency of at least 80 throughout the CSC and a range of 50% across the sub-corpora which comprise it. Finally, the *Collocates* tool was useful to identify the frequency of collocational activity for a specific token to be examined. It could do this by counting the frequency with which other tokens to a number of places left or right of the original search token appearred.

There were two separate stages to the process of identifying multi-word units within the CSC. They will be outlined in the following sub-sections.

5.5.1 Hyphenated and compound words in the CSWL

AntWordProfiler (Anthony, 2008) is not designed to handle non-alphabetic characters. This is why all such items were removed during editing of the CSC (section 4.3). However, this had to be accounted for as the possibility remained that words from the CSWL which were previously hyphenated were lost at this stage. Similarly, a compound word could have been written as a multi-word unit. This seemed likely as many of the headwords in the CSWL were either compound words, could be hyphenated or contained a known affix, such as *workload*, *multilevel* or *email*. In order to identify this, the headword was stripped of its affix or divided into its constituent parts and run through the *Collocates* tool. If the type which had been removed appeared on the list of collocates, its frequency was noted. This would have effectively contributed to the overall coverage provided by the CSWL within the results so far. Table 5.9 shows some of the the results of this query.

Headword	**Constituent parts**	**Extra frequency of occurrence**	**Percentage of headword frequency**
coefficient	co + efficient	1	0.68%
dataset	data + set	214	30.62%
email	e + mail	21	23.86%
hardware	hard + ware	6	0.48%
healthcare	health + care	57	50.00%
internet	inter + net	3	0.55%
multilevel	multi + level	15	14.42%
offline	off + line	18	8.49%
online	on + line	96	11.58%
update	up + date	2	0.20%

Table 5.9 Examples of hyphenated or multi-word units in the CSWL

Overall, a total of 46 headwords from the CSWL also existed within the CSC as part of a multi-word unit or hyphenated word. This increased the total frequency of these words by an additional 887 occurrences. Their grammatical usage was accounted for manually during the counting process. This increase in the number of occurrences of the CSWL headwords would have provided an extra

0.05% coverage of the CSC. Unfortunately, none of the software available for profiling frequency data from corpora is capable of counting hyphenated words or multi-word units.

5.5.2 Multi-word units outside of the CSWL

Having identified that 46 words from the CSWL also existed as either hyphenated words or as separate multi-word units, the next step was to count how many phrases outside of the CSWL existed within the CSC. In order to do this, AntConc (Anthony, 2002) was again employed. Firstly, an *ngram* analysis of the CSC was run, looking for clusters of between 2 and 5 tokens which regularly co-occurred. A total of 5,562 such expressions with a frequency of 80 or greater were detected. These were manually compared to the ODOC and, after evaluation, a total of only 34 remained. As Ward (2007) discovered in his research, the majority of these were noun phrases. These were run through the *Concordance Plot* tool to ensure range was at least equal to 50% and 28 phrases remained. Multi-word units like *phase change* and *machine translation* were lost at this stage. This process ensured that any expressions to be added to the CSMWL matched the criteria Coxhead (2000) established for the selection of words for the AWL. It also ensured that none of the words from the CSWL existed only as part of a multi-word unit, as the same criteria were used (section 4.7). If a word from the CSWL occurred with the same frequency as it did in a phrase, it would have appeared in this list.

These multi-word units were then expanded to level 6 (Bauer & Nation, 1993) as some were clearly plural forms of other phrases such as *operating systems* and *social networks*. This left 23 separate multi-word units which met the necessary criteria for inclusion in the CSMWL. It was not possible to achieve specialised occurrence (Coxhead, 2000) as the tokens which comprised occurred within other word lists. The CSMWL may be found in Appendix B.

An interesting aspect of the CSMWL is that if it were added to the CSWL, then it would replace one of the entries. The headword *garbage* is taken from the CSWL and it has a frequency in the CSC of 169. The headword *garbage collection* is taken from the CSMWL. It has a frequency in the CSC of 140. This means that there are only 29 occurrences of the word *garbage* which were not followed by the word *collection* which would have been insufficient frequency to have qualified for the CSWL originally. Also, *dataset* exists within the CSWL and *data set* within the CSMWL without violating any selection criteria. These highlight the importance of research into multi-word units in technical word lists and their overlap with single word technical lists.

All of the 23 items on the CSWL consisted of 2 or 3 token multi-word units only. This gave a total of 42 separate types as some were repeated, such as in *lower bound* and *upper bound*. These were run in AntWordProfiler (2008) against the GSL, AWL and CSWL (Table 5.10).

Word List	Number of types in the CSMWL	Percentage coverage of CSMWL
GSL1	17	40.48%
GSL2	11	26.19%
AWL	8	19.05%
CSWL	6	14.28%
Total	**42**	**100.00%**

Table 5.10 Distribution of types in the CSMWL against the other word lists

This finding shows that all of the lexical types from the CSMWL are present in the main word lists investigated so far in this study. These were mostly derived from high frequency vocabulary (66.67%). This supports Martinez and Schmitt's (2012) conclusion that multi-word units may be deceptively transparent as learners may know all the words from the GSL, AWL and CSWL but not understand their meaning when combined in certain ways. However, this may be taken under advisement as these have definite technical meanings which a student would have to learn to recognise anyway.

Nonetheless, the clear presence of multi-word units with sufficient frequency and range within the CSC does support the contention that there are multi-word units within computer science literature in answer to the final research question of this study (chapter 3).

6. Discussion

6.1 Introduction

The literature review (chapter 2) highlighted some omissions in current vocabulary research which this study hoped to redress. Primary amongst these was the lack of any research into the subject specific lexis of computer science. Whilst other researchers (e.g. Coxhead & Hirsch, 2007; Wang et al, 2008; Ward, 1999) have covered other scientific disciplines, none have considered this increasingly popular and influential discipline. This provided the motivation for this paper and the development of the CSWL.

Other gaps in the current state of vocabulary research were also noted. They provided the bases for the research questions posed (chapter 3). Once a methodology had been established (chapter 4), the results from the experimental data provided a mix of both conclusive and tentative answers to these questions (chapter 5). These results will now be considered in a larger context, related where possible to previous research in this field and highlighting any problematic data or unexpected results.

6.2 Coverage of the GSL and AWL

The first research question (chapter 3) asked how much coverage the GSL and AWL would provide within a computer science corpus. Given a balanced, representative, sufficiently sized corpus, these results could then be extrapolated to say something about the coverage of academic computer science literature in a more general fashion.

The GSL accounted for 76.43% of all tokens within the CSC and 77.45% in the test computer science corpus. Compared to the coverage it provided on similar studies (Table 5.2), this was very low indeed. Konstantakis (2007) achieved 85.72% coverage and Aichah (2012) 81.79%. It might even be considered comparable to the 76.1% achieved by Coxhead (2000). However, when single letters (other than *A* and *I*) were removed from the GSL1 word list, the overall coverage of the GSL dropped to 69.86% in the CSC. In fact, a total of 152 word families from the GSL, which are meant to represent the most common words in English, were not found within the CSC. These results demonstrate 2 issues. Firstly, performing a corpus analysis of computer science literature is difficult

as a significant proportion of the data is not written in English, but in the languages of mathematics and computer programming. These code switches are frequently performed mid-sentence and so are not clearly delineated from the English surrounding them. Secondly, and despite the arguments of Gilner (2011), it may be that the GSL it not as efficient a word list for teaching students of the computer science discipline as it may be for those of other academic subjects.

These results might suggest that Ward's (1999) decision to create a single word list was a preferable methodology, until the coverage of the AWL is considered. The AWL performed very well against all technical corpora in this study. Coxhead (2011) illustrated how on average the AWL provides 10% coverage across the technical word lists studies available for analysis. With regards to the CSC, the AWL accounted for 12.79% of all tokens and for a comparable 12.28% within the test computer science corpus. Only 2 word families of the AWL were not present in the CSC. It was not entirely unexpected that the AWL should perform so well in computer science literature. It contains many polysemic words (Wang & Nation, 2004) with a computer science bias, which had very high representation within the corpus data of this subject. The headwords *data*, *process*, *section*, *compute* and *network* were the 5 most frequent AWL words in the CSC. Any further studies into computer science vocabulary should incorporate the AWL.

6.3 Contents and coverage of the CSWL

The methodology discussed in this paper (chapter 4) outlined the way in which the CSWL was built. It was decided during the compilation of the CSWL that no subjective evaluations were to be made about the headwords chosen for inclusion. As long as they met the necessary criteria (chapter 4.6) that was sufficient. This resulted in the incorporation of many words which might be considered general purpose or even related to other academic disciplines. The headword *afore* is present in the CSWL with the only member of the family being *aforementioned*. This is a general purpose discourse marker which has relatively frequent use within the literature of the computer science discipline, but has no particular reason to be prevalent within it. *Healthcare* was another unexpected headword within the CSWL and is normally associated with other subjects, but it is there because it met the criteria.

The CSWL contained many words which might otherwise be hyphenated or even written as separate words (section 5.5.1). This resulted in the inclusion of words like *dataset* into the CSWL. *Data* is a word on the AWL but was allowed for inclusion within the CSWL as it occurred

frequently enough and with sufficient range as a compound noun to qualify. The criterion of specialist occurrence is not broken here as *dataset* is certainly not within level 6 expansion of the word *data* and must be treated as a separate headword in it own right. The same judgement applied to *database*, *timestamp* and others. Analysis showed that no words included within the CSWL existed more frequently in their hyphenated or individual constituent forms (Table 5.9).

The coverage obtained by the CSWL varied. Whilst it covered 6.00% of the CSC, its coverage dropped to 4.68% within the test computer science corpus. This is a relatively high amount of coverage per headword compared to other studies (Table 6.1).

	CSWL	**LWL (Aichah, 2012)**	**Science Word List (Coxhead & Hirsch, 2007)**
Headwords	433	373	318
Coverage	6.00%	3.86%	3.79%
Coverage per headword	**0.014**	**0.010**	**0.012**

Table 6.1 Coverage per headword

Overall, it was not possible to achieve the lexical threshold for sufficient understanding of a text of 95% (Laufer, 1989) when the CSWL was employed in the test computer science corpus, the CSCPC and within 2 of the sub-discipline corpora. Although a reduction in coverage of the other word lists was a contributory factor to some of these unsatisfactory results, it might still be said that the CSWL did not consistently provide sufficient supplemental coverage to the GSL and AWL. Certainly in terms of coverage alone, the CSWL was a qualified, partial success in answer to the second research question (chapter 3).

6.4 Efficiency of the GSL/AWL/CSWL against the BNC frequency bands

The third research question (chapter 3) asked if it was more efficient to learn the combined GSL, AWL and CSWL (2,992 headwords) than it would be to learn however many frequency bands from the BNC (Nation & Heatley, 2002) which it would take to achieve the same coverage of the CSC.

Analysis showed that 95% coverage could only be achieved by the first 10k BNC frequency bands. This means that 10,000 words would have to be learned by a L2 learner of English studying

computer science in place of 2,992 suggested by the other word lists, a ratio of 3.34:1 (BNC: GSL/AWL/CSWL). As this is over 3 times as many words, it is clearly not as efficient a practice and this has important pedagogical implications. Aichah (2012) found that his LC could be covered by only the first 4k BNC frequency bands, representing a ratio of 1.36:1. So it could be argued that the combined GSL, AWL and CSWL represent a particularly efficient vocabulary set for the instruction of the intended benefactors of this study.

Equally important was the demonstration that it was impossible to achieve the 98% lexical threshold (Hirsch & Nation, 1992; Hu and Nation, 2000; Schmitt, Jiang & Grabe, 2011) using all 14k frequency bands of the BNC or even with the 25k frequency bands of the BNC/COCA (Nation & Heatley, 2002). It certainly raises the notion that lexical thresholds reach a natural ceiling of around 95% in technical word lists, which is the standard Coxhead (2000) established for the AWL. If a technical text contains a higher proportion of technical lexis than non-technical texts (Hirsch, 2004), and this technical lexis is distributed through the mid-frequency to low-frequency bands (Laufer & Ravenhorst-Kalovski, 2010), then a greater number of headwords per token are required to provide coverage in a technical text than in a non-technical one and this increases exponentially with the density of technical lexis. None of the studies which postulated a 98% lexical threshold (Hirsch & Nation, 1992; Hu and Nation, 2000; Schmitt, Jiang & Grabe, 2011) used technical corpora to calculate this figure so this value should be considered as inappropriate to any technical corpus study.

6.5 Technicality and distribution of the CSWL

The fourth research question of this study (chapter 3) asked if the CSWL was a technical word list, but there are many ways in which the technicality of a word may be determined (Chung & Nation, 2003). There appears to be an overlap between categories of words which may be considered technical or academic and they may be subject dependent in meaning as well as frequency (Nation, 2001a). This means that a single process to identify the technicality of the CSWL was infeasible. Instead, the question had to be approached from a number of different directions as suggested by previous research (e.g. Chung & Nation; Coxhead, 2000).

Some research (Laufer & Ravenhorst-Kalovski, 2010; Schmitt & Schmitt, 2012) had shown the distribution of academic and thus technical lexis throughout the mid-frequency vocabulary range. Schmitt & Schmitt (2012) noted that 64.3% of the headwords of the AWL occurred within the high-

frequency rang (i.e. the first 3k word frequency bands). This study, using the BNC 14k frequency bands and the AWL by Nation and Heatley (2002) found this figure to be 66.58%. In addition, 31% of the AWL occurred within the mid-frequency range (3k-9k frequency bands) and the remaining 2.42% within the low-frequency range (10k+ frequency bands). The CSWL, in contrast, had 47.24% of its headwords in this mid-frequency range where technical/academic lexis might be expected. There was a fairly even distribution of the rest of its headwords between the high and low-frequency bands of 24.01% and 28.75% respectively (Figure 5.1). The distribution of headwords within the CSWL strongly supported its technicality.

Checking the CSWL against a second, test computer science corpus was a standard positive control on the technicality of its headwords, recommended by Coxhead (2000). If the CSWL was representative of computer science vocabulary and thus technical in nature, it would be expected to have a similar level of coverage over any corpus of subject specific literature. The result was inconclusive. The study was unable to obtain the 95% lexical threshold required for sufficient understanding (Laufer, 1989) of the test computer science corpus using the GSL, AWL and CSWL alone. Only 94.41% coverage was obtained and the CSWL's performance dropped from 6% (against the CSC) to only 4.88%. However, profiling the headwords of the CSWL using the *Concordance* tool from AntConc (Anthony, 2002) showed that this problem could be due to a difference in spelling conventions. *E-mail* was 5 times more frequent than *email*, in the test corpus, but hyphenated words or compound words split into their constituent parts were not included into the coverage count.

Testing the CSWL against a fiction corpus was a negative control (Coxhead, 2000). It was not expected that it would provide as much coverage of this genre of literature and thus show more evidence for the technicality of its headwords. The results here strongly supported this argument. 86.15% of all headwords of the CSWL were demonstrated to be technical, according to Coxhead's (2000) definition i.e. they were either not present in the fiction corpus or 4 times or more frequent in the technical corpus than in the fiction one.

Finally, in addition to the corpus analytical approaches, another method of identifying technical vocabulary by comparison to a technical dictionary was recommended by Chung and Nation (2003). On comparison with the ODOC, 70.2% of all headwords (or their different lexical types) in the CSWL were discovered to either have a main entry in the ODOC or were found as part of a multi-word entry. Aichah (2012) found only 58.2% technicality in his LWL in this way and

Konstantakis (2010) found only 54%. Whilst Konstantakis (2010) only looked for headword main-entries, it is still evident that the technicality of the CSWL has been strongly supported by this process.

In fact, the overwhelming majority of evidence gained by these 4 experiments demonstrated the highly technical nature of the CSWL, in so far as lexical technicality is currently defined (e.g. Chung & Nation, 2003l; Coxhead, 2000; Schmitt & Schmitt, 2012).

6.6 Multi-word units in the CSC

The final question of this study (chapter 3) asked if there were any multi-word units in computer science. In order to answer this question, it was necessary to detect any multi-word units in the CSC so a result could be extrapolated from that. Again, this proved to be a complicated process as many of the headwords of the CSWL demonstrated a number of different spelling conventions which could lead to them existing in a compound form (as in the CSWL), a hyphenated or even multi-word unit form. Also, for these multi-word units to be accepted as part of computer science discourse and to be added to the CSMWL, they had to exhibit the same criteria of range and frequency as demanded of words in the CSWL.

The first process was to ensure that no entries on the CSWL existed only as part of a multi-word unit and once that was determined, to check the frequency with which they occurred with other spelling conventions which the software would read as multi-word units. No evidence for the headwords of the CSWL existing only as part of a multi-word unit was found. Some evidence showed that words on the CSMWL would replace words on the CSWL, as they were found most commonly as a part of formulaic language e.g. *garbage* and *garbage collection*. It was even found that words could co-exist on both word lists, such as *dataset* and *data set*. A total of 46 headwords of the CSWL were found to have variations in spelling conventions leading to an effective under-representation of the CSWL in the research. Finally, with all the cluster data collected from the CSC, a final CSMWL of 23 items was compiled.

Very little research has been carried out into the presence of multi-word units in technical genres. It is a relatively new area of research interest which might explain this deficit. Martinez and Schmitt (2012) produced a PHRASE list of high-frequency formulaic language only in this last year. Durrant (2009) looked at formulaic language in the AWL and found a range of lexical and

grammatical collocations and this helped determine that a technical multi-word list should only exist of subject specific phrases. This was the methodology used by Aichah (2012) and also adopted for this study. The results were similar to those obtained by Ward (2007) in that every item on the CSMWL consisted of noun phrases, but this was to be expected given the criterion of selection of these items by comparison with a technical dictionary. In this way, the CSMWL only proved that items in the ODOC occured within the CSC and hence computer science literature generally. This was an expected result and only demonstrates a low coverage value for multi-word technical units in a technical corpus. Such a methodology could not hope to demonstrate the ubiquity of formulaic language as discovered in other research (Martinez & Murphy, 2011; Martinez & Schmitt, 2012) as it dealt with infrequent combinations of infrequent vocabulary.

7. Conclusion

The purpose of this study was to identify the technical vocabulary used within the discipline of computer science and extract this to form a word list which would help L2 learners of English who are studying this subject in an English medium. This resulted in 2 separate technical word lists: the CSWL which consisted of 433 headwords, and the CSMWL which contained 23 computer science specific formulaic expressions. When combined with the GSL and AWL they provided over 95% coverage which has become a standard target for studies of this kind. In this way, the hypothesis that a CSWL could help L2 learners of English studying computer science in the UK was confirmed. Moreover, analysis of the CSC demonstrated the density of technical lexis in computer science literature which highlights the insufficiency of general English instruction for such students.

8. Limitations and suggestions for future research

An increase in sample size improves the veracity of any claims drawn from it. This applies to corpus analysis such that a larger corpus is generally considered as being preferable. However, when external criteria are applied to a corpus in the building of a word list derived from it, then the non-linear relationship between word frequency and corpus size needs to be fully considered. This was a problem which restricted the size of the technical corpus in this study to that of other studies (Coxhead, 2000) and is an area where further research is necessary.

Finally, there is a methodological and pedagogical dilemma involved in the inclusion of multi-word units in technical vocabulary studies. Technical formulaic language is more likely to require instruction by a subject specialist. It has a meaning separate and distinct from the sum of the individual words from which it is comprised and so is a part of language which may fall outside of the remit of ESP teaching. This is certainly the case when a dictionary is used as a means of identifying such language. An interesting corollary is that previous research into technical word lists could be error-checked. By comparing the incidence of headwords as part of a collocation only, their validity as a headword on a technical word list could be verified. This also suggests that further research into multi-word units could focus on subject independent corpora as further work is needed into the presence of formulaic language in both general and academic use.

APPENDIX A: The Computer Science Word List (CSWL)

Headwords of the CSWL in alphabetical order			
accelerate	activate	acyclic	adversary
affine	afore	algebra	algorithm
align	alphabet	amortise	annotate
anomaly	anonymous	arc	architecture
arithmetic	array	artifact	asynchronous
atom	audio	augment	authentic
authorise	automaton	autonomic	auxiliary
avail	axis	backup	bandwidth
barrier	baseline	batch	battery
bayesian	benchmark	binary	binomial
bitmap	boolean	bottleneck	breakdown
browse	budget	buffer	bug
byte	cache	calculus	calibrate
candidate	canonical	capture	cell
cellular	chip	chunk	churn
circuit	click	client	cluster
coefficient	cognitive	collaborate	collision
column	compact	compiler	compress
compromise	concrete	concurrency	configure
congest	conjecture	conjunction	consecutive
contend	contiguous	contour	contraction
converge	convex	convolute	corollary
corpus	correlate	corrupt	counter
cryptography	customise	database	dataset
deadline	debug	decentralize	decode
decompose	decrypt	dedicate	default
defect	degrade	delete	dense
departure	dependency	depict	deploy
descriptor	destination	diagnosis	diagonal
diagram	diameter	differential	digital
disc	discard	disclosure	disseminate
download	drawback	dual	efficacy
electronic	email	embed	embody

emergency emotion emulate encode
encrypt endpoint enterprise entropy
epoch equilibrium ethernet execute
existential exit exponential faculty
fake feasible feedback fetch
filter firewall footprint forum
fraction fragment functionality fuse
fuzz gadget garbage genetic
genre geometry gesture ghost
gossip gradient granule graph
grid hammed handshake hardware
hash header healthcare heterogeneous
heuristic histogram homogeneous hop
horizontal huge hybrid icon
identifier incoming increment incur
indirect inductive infect inject
inspire install instantiate integer
interface internet interoperability intersect
interview intrusion intuition inverse
iterate jitter kernel keyboard
laboratory latent lattice layout
leak legitimate lemma lever
linear literal locality logarithm
lookup loop magnet magnitude
malicious mask mathematics matrix
maximal median merge mesh
metadata metric microprocessor mobile
modal module monotone multicore
multilevel multimedia multimode multithread
mutate naive navigate negligible
neural node notate novel
null offline offload online
ontology opt optic optimum
optimise orthogonal outgoing outlying
outperform overflow overhead overlay
overview packet pairwise partition

password payload payoff peak
peer penalty periphery personalise
pervasive phrase pilot pipeline
pixel planar platform plot
polynomial port posterior predecessor
predicate prefix primitive privacy
probe processor profile prominent
prone propagate proposition prototype
proximity prune pulse quadratic
quantify query queue radius
recall reconfigure rectangle recursive
redundant regress remote render
replicate repository residue resilience
retrieve robust rout routine
rotate runtime scan schema
score script segment semantic
sensor serial session setup
sibling simulations simultaneous sketch
skip slack slot smart
smartphone snapshot software sophisticated
soundly spam span sparse
spatial spectre spectrum speedup
stack static stationary stochastic
storage subjective subscribe substrate
suffix suite superior supervise
swap switch symmetry synchronize
syntactic syntax synthesis tablet
tag template temporal testbed
texture theorem thermal threshold
throughput tier timestamp timing
token tolerant topology traffic
transact transient transitive transparent
traverse triple trivial trustworthy
tuple ubiquity update upload
usage vector velocity verify
versus vertex vertice vertical

vice	victim	video	virtualize
vocabulary	volt	vulnerable	wavelet
web	wireless	workflow	workload
workstation			

Appendix B: The Computer Science Multi-Word List (CSMWL)

Headwords of the CSMWL in alphabetical order	
control flow graph	data flow
data mining	data set
data structure	data transfer
lower bound	flash memory
execution time	garbage collection
machine learning	operating system
polynomial time	response time
scratch pad	search engine
social network	software development
software engineer	steady state
upper bound	user interface
virtual machine	

Appendix C: CSCPC Bibliography

Computer systems organisation

Abbasi, H., Eisenhauer, G., Wolf, M., Schwan, K., & Klasky, S. (2011, June). Just in time: adding value to the IO pipelines of high performance applications with JITStaging. In *Proceedings of the 20th international symposium on High performance distributed computing* (pp. 27-36). ACM.

Afek, Y., Morrison, A., & Wertheim, G. (2011, June). From bounded to unbounded concurrency objects and back. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 119-128). ACM.

Al-Otoom, M., Forbes, E., & Rotenberg, E. (2010, May). EXACT: Explicit dynamic-branch prediction with active updates. In *Proceedings of the 7th ACM international conference on Computing frontiers* (pp. 165-176). ACM.

Avin, C., Borokhovich, M., Censor-Hillel, K., & Lotker, Z. (2011, June). Order optimal information spreading using algebraic gossip. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 363-372). ACM.

Chen, M. H., & Chou, P. H. (2010, October). TeleScribe: a scalable, resumable wireless reprogramming approach. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 139-148). ACM.

Chokshi, R., Berezowski, K. S., Shrivastava, A., & Piestrak, S. J. (2009, October). Exploiting residue number system for power-efficient digital signal processing in embedded processors. In *Proceedings of the 2009 international conference on Compilers, architecture, and synthesis for embedded systems* (pp. 19-28). ACM.

Dobre, D., Guerraoui, R., Majuntke, M., Suri, N., & Vukolić, M. (2011, June). The complexity of robust atomic storage. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 59-68). ACM.

Feng, T. H., Lee, E. A., & Shruben, L. W. (2010, October). Ptera: an event-oriented model of computation for heterogeneous systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 219-228). ACM.

Gray, I., & Audsley, N. C. (2011, April). Targeting complex embedded architectures by combining the multicore communications API (mcapi) with compile-time virtualisation. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 51-60). ACM.

Hahn, J., & Chou, P. H. (2010, October). Nucleos: a runtime system for ultra-compact wireless sensor nodes. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 149-158). ACM.

Hashmi, A., Nere, A., Thomas, J. J., & Lipasti, M. (2012). A case for neuromorphic ISAs. *ACM SIGPLAN Notices*, *47*(4), 145-158.

Hilton, A., & Roth, A. (2009, June). Decoupled store completion/silent deterministic replay: Enabling scalable data memory for CPR/CFP processors. In *ACM SIGARCH Computer*

Architecture News (Vol. 37, No. 3, pp. 245-254). ACM.

Kessler, C. W., & Keller, J. (2009). Optimized on-chip pipelining of memory-intensive computations on the Cell BE. *ACM SIGARCH Computer Architecture News*, *36*(5), 36-45.

Lee, J., Shin, I., & Easwaran, A. (2010, October). Online robust optimization framework for qos guarantees in distributed soft real-time systems. In*Proceedings of the tenth ACM international conference on Embedded software* (pp. 89-98). ACM.

Mittal, R., Kansal, A., & Chandra, R. (2012, August). Empowering developers to estimate app energy consumption. In *Proceedings of the 18th annual international conference on Mobile computing and networking* (pp. 317-328). ACM.

Santinelli, L., Marinoni, M., Prosperi, F., Esposito, F., Franchino, G., & Buttazzo, G. (2010, October). Energy-aware packet and task co-scheduling for embedded systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 279-288). ACM.

Sarkar, A., Mueller, F., & Ramaprasad, H. (2011, April). Predictable task migration for locked caches in multi-core systems. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 131-140). ACM.

Stoimenov, N., Thiele, L., Santinelli, L., & Buttazzo, G. (2010, October). Resource adaptations with servers for hard real-time systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 269-278). ACM.

Suri, T., & Aggarwal, A. (2009, May). Improving performance of simple cores by exploiting loop-level parallelism through value prediction and reconfiguration. In *Proceedings of the 6th ACM conference on Computing frontiers* (pp. 151-160). ACM.

Taly, A., & Tiwari, A. (2010, October). Switching logic synthesis for reachability. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 19-28). ACM.

Vaquero, L. M., Rodero-Merino, L., & Buyya, R. (2011). Dynamically scaling applications in the cloud. *ACM SIGCOMM Computer Communication Review*, *41*(1), 45-52.

Wang, H., Koren, I., & Krishna, C. M. (2008, October). An adaptive resource partitioning algorithm for SMT processors. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques* (pp. 230-239). ACM.

Westermann, D., Happe, J., Krebs, R., & Farahbod, R. (2012, September). Automated inference of goal-oriented performance prediction functions. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering* (pp. 190-199). ACM.

Yao, G., & Buttazzo, G. (2010, October). Reducing stack with intra-task threshold priorities in real-time systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 109-118). ACM.

Computing methodologies

Ahmad, J. J., Li, S., Amer, I., & Mattavelli, M. (2011, September). Building multimedia security applications in the MPEG Reconfigurable Video Coding (RVC) framework. In *Proceedings of the*

thirteenth ACM multimedia workshop on Multimedia and security (pp. 121-130). ACM.

Albert, E., Arenas, P., Genaim, S., & Zanardini, D. (2011). Task-level analysis for a language with async/finish parallelism. *ACM SIGPLAN Notices*, *46*(5), 21-30.

Alistarh, D., Aspnes, J., Censor-Hillel, K., Gilbert, S., & Zadimoghaddam, M. (2011, June). Optimal-time adaptive strong renaming, with applications to counting. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 239-248). ACM.

Aronis, S., Papaspyrou, N., Roukounaki, K., Sagonas, K., Tsiouris, Y., & Venetis, I. E. (2012, September). A scalability benchmark suite for Erlang/OTP. In *Proceedings of the eleventh ACM SIGPLAN workshop on Erlang workshop* (pp. 33-42). ACM.

Aspnes, J., Attiya, H., Censor-Hillel, K., & Ellen, F. (2012, July). Faster than optimal snapshots (for a while): preliminary version. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing* (pp. 375-384). ACM.

Bonakdarpour, B., Bozga, M., Jaber, M., Quilbeuf, J., & Sifakis, J. (2010, October). From high-level component-based models to distributed implementations. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 209-218). ACM.

Burdis, J. M., & Kogan, I. A. (2012, June). Object-image correspondence for curves under central and parallel projections. In *Proceedings of the 2012 symposuim on Computational Geometry* (pp. 373-382). ACM.

Bushkov, V., Guerraoui, R., & Kapałka, M. (2012, July). On the liveness of transactional memory. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing* (pp. 9-18). ACM.

Chafi, H., DeVito, Z., Moors, A., Rompf, T., Sujeeth, A. K., Hanrahan, P., ... & Olukotun, K. (2010, October). Language virtualization for heterogeneous parallel computing. In *ACM Sigplan Notices* (Vol. 45, No. 10, pp. 835-847). ACM.

Delporte-Gallet, C., Fauconnier, H., Guerraoui, R., Kermarrec, A. M., & Ruppert, E. (2011, June). Byzantine agreement with homonyms. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 21-30). ACM.

Dragojević, A., Herlihy, M., Lev, Y., & Moir, M. (2011, June). On the power of hardware transactional memory to simplify memory management. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 99-108). ACM.

Foltzer, A., Kulkarni, A., Swords, R., Sasidharan, S., Jiang, E., & Newton, R. (2012, September). A meta-scheduler for the par-monad: composable scheduling for the heterogeneous cloud. In *Proceedings of the 17th ACM SIGPLAN international conference on Functional programming* (pp. 235-246). ACM.

Ha, O. K., Kuh, I. B., Tchamgoue, G. M., & Jun, Y. K. (2012, July). On-the-fly detection of data races in OpenMP programs. In *Proceedings of the 2012 Workshop on Parallel and Distributed Systems: Testing, Analysis, and Debugging* (pp. 1-10). ACM.

Kaspar, M., Parsad, N. M., & Silverstein, J. C. (2010, November). CoWebViz: interactive collaborative sharing of 3D stereoscopic visualization among browsers with no added software. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 809-816). ACM.

Liu, J., & Xia, S. (2010, November). Effective epidemic control via strategic vaccine deployment: a systematic approach. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 91-99). ACM.

Lou, Y., Caruana, R., & Gehrke, J. (2012, August). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150-158). ACM.

Marquet, K., & Moy, M. (2010, October). PinaVM: a SystemC front-end based on an executable intermediate representation. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 79-88). ACM.

Matsushima, S., Vishwanathan, S. V. N., & Smola, A. J. (2012, August). Linear support vector machines via dual cached loops. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 177-185). ACM.

Neerbek, J. (2012, August). Message-driven FP-growth. In *Proceedings of the WICSA/ECSA 2012 Companion Volume* (pp. 29-36). ACM.

Ng, M. K., Wu, Q., & Ye, Y. (2012, August). Co-transfer learning via joint transition probability graph based method. In *Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining* (pp. 1-9). ACM.

Phan, L. T. X., Schneider, R., Chakraborty, S., & Lee, I. (2010, October). Modeling buffers with data refresh semantics in automotive architectures. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 119-128). ACM.

Richards, D. (2012, July). Agent-based museum and tour guides: applying the state of the art. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System* (p. 15). ACM.

Stevenson, J. P., Firoozshahian, A., Solomatnikov, A., Horowitz, M., & Cheriton, D. (2012, June). Sparse matrix-vector multiply on the HICAMP architecture. In *Proceedings of the 26th ACM international conference on Supercomputing* (pp. 195-204). ACM.

Toscano, L., D'Angelo, G., & Marzolla, M. (2012, September). Parallel discrete event simulation with Erlang. In *Proceedings of the 1st ACM SIGPLAN workshop on Functional high-performance computing* (pp. 83-92). ACM.

Wang, J., Yang, Z., Jin, L., Deng, J., & Chen, F. (2010, September). Adaptive surface reconstruction based on implicit PHT-splines. In *Proceedings of the 14th ACM Symposium on Solid and Physical Modeling* (pp. 101-110). ACM.

Hardware

Agrawal, S., Athota, K., Bhatotia, P., Goyal, P., Krisha, P., Ruchandan, K., ... & Yu, F. (2011). Session reports for SIGCOMM 2010. *Computer Communication Review*, *41*(1), 66-83.

Albarghouthi, A., Kumar, R., Nori, A. V., & Rajamani, S. K. (2012). Parallelizing top-down interprocedural analyses. *ACM SIGPLAN Notices*, *47*(6), 217-228.

Braun, A., & Hamisu, P. (2011, May). Designing a multi-purpose capacitive proximity sensing input device. In *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments* (p. 15). ACM.

Byers, J. W., Mogul, J. C., Adib, F., Aikat, J., Chasaki, D., Chen, M. H., ... & Yang, C. Y. (2012). Report on the SIGCOMM 2011 Conference. *SIGCOMM-Computer Communication Review*, *42*(1), 80.

Cho, K., We, K. S., Lee, C. G., & Kim, K. (2010, October). Using NAND flash memory for executing large volume real-time programs in automotive embedded systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 159-168). ACM.

Chou, C. H., Pong, F., & Tzeng, N. F. (2012, February). Speedy FPGA-based packet classifiers with low on-chip memory requirements. In *Proceedings of the ACM/SIGDA international symposium on Field Programmable Gate Arrays* (pp. 11-20). ACM.

Cullmann, C. (2011, April). Cache persistence analysis: a novel approachtheory and practice. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 121-130). ACM.

Golab, W. (2011, June). A complexity separation between the cache-coherent and distributed shared memory models. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 109-118). ACM.

Grover, A. S., Calteaux, K., Barnard, E., & van Huyssteen, G. (2012, March). A voice service for user feedback on school meals. In *Proceedings of the 2nd ACM Symposium on Computing for Development* (p. 13). ACM.

Gürkök, H., Hakvoort, G., & Poel, M. (2011, November). Modality switching and performance in a thought and speech controlled computer game. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 41-48). ACM.

Jin, S., Kim, J., Kim, J., Huh, J., & Maeng, S. (2011, March). Sector log: fine-grained storage management for solid state drives. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (pp. 360-367). ACM.

Koradia, Z., & Seth, A. (2012, March). Phonepeti: exploring the role of an answering machine system in a community radio station in india. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (pp. 278-288). ACM.

Kjærgaard, M. B., Wirz, M., Roggen, D., & Tröster, G. (2012, September). Detecting pedestrian flocks by fusion of multi-modal sensors in mobile phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 240-249). ACM.

Nageswaran, U. B., & Chilambuchelvan, A. (2012, August). High speed VLSI implementation of

lifting based DWT. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 1104-1110). ACM.

Nageswaran, U. B., & Chilambuchelvan, A. (2012, August). VLSI architectures for lifting based DWT: a detailed survey. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 207-214). ACM.

Owens, K. V., Nagorski, A., Myers, S., Wezeman, J., & Pettigrew, R. (2010, October). Print centers: navigating the sea of ink. In *Proceedings of the 38th annual fall conference on SIGUCCS* (pp. 51-60). ACM.

Patel, N., Shah, K., Savani, K., Klemmer, S. R., Dave, P., & Parikh, T. S. (2012, March). Power to the peers: authority of source effects for a voice-based agricultural information service in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (pp. 169-178). ACM.

Phung, P. H., & Desmet, L. (2012, June). A two-tier sandbox architecture for untrusted javascript. In *Proceedings of the Workshop on JavaScript Tools* (pp. 1-10). ACM.

Qadeer, W., Hameed, R., Shacham, O., Venkatesan, P., Kozyrakis, C., & Horowitz, M. A. (2013, June). Convolution engine: balancing efficiency & flexibility in specialized computing. In *Proceedings of the 40th Annual International Symposium on Computer Architecture* (pp. 24-35). ACM.

Ryu, M., Kim, H., & Ramachandran, U. (2011, February). Impact of flash memory on video-on-demand storage: analysis of tradeoffs. In *Proceedings of the second annual ACM conference on Multimedia systems* (pp. 175-186). ACM.

Yan, S., Wu, C., Dai, W., Ghanem, M., & Guo, Y. (2012, August). Environmental monitoring via compressive sensing. In *Proceedings of the Sixth International Workshop on Knowledge Discovery from Sensor Data* (pp. 61-68). ACM.

Zhang, Q., Jin, H., Liao, X., Li, D., & Deng, W. (2012, February). FIOS: a flexible virtualized I/O subsystem to alleviate interference among virtual machines. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication* (p. 61). ACM.

Zhao, H., Jang, O., Ding, W., Zhang, Y., Kandemir, M., & Irwin, M. J. (2012, June). A hybrid NoC design for cache coherence optimization for chip multiprocessors. In *Proceedings of the 49th Annual Design Automation Conference* (pp. 834-842). ACM.

Zúquete, A., & Frade, C. (2012). A new location layer for the TCP/IP protocol stack. *ACM SIGCOMM Computer Communication Review*, *42*(2), 16-27.

Human-centred computing

Alencar, A. B., Börner, K., Paulovich, F. V., & de Oliveira, M. C. F. (2012, March). Time-aware visualization of document collections. In*Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 997-1004). ACM.

Alves, V., & Roque, L. (2011, November). A deck for sound design in games: enhancements based

on a design exercise. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (p. 34). ACM.

Arnold, K. C., & Lieberman, H. (2010, October). Managing ambiguity in programming by finding unambiguous examples. In *ACM Sigplan Notices* (Vol. 45, No. 10, pp. 877-884). ACM.

Bakker, S., van den Hoven, E., Eggen, B., & Overbeeke, K. (2012, February). Exploring peripheral interaction design for primary school teachers. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction* (pp. 245-252). ACM.

Blouin, A., Morin, B., Beaudoux, O., Nain, G., Albers, P., & Jézéquel, J. M. (2011, June). Combining aspect-oriented modeling with property-based reasoning to improve user interface adaptation. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 85-94). ACM.

Eickhoff, C., Dekker, P., & de Vries, A. P. (2012, August). Supporting children's web search in school environments. In *Proceedings of the 4th Information Interaction in Context Symposium* (pp. 129-137). ACM.

Fleury, A. (2012, August). Drawing and acting as user experience research tools. In *Proceedings of the 10th asia pacific conference on Computer human interaction* (pp. 269-278). ACM.

Hammond, T., & Paulson, B. (2011). Recognizing sketched multistroke primitives. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *1*(1), 4.

Hoskinson, R., Stoeber, B., Heidrich, W., & Fels, S. (2010, December). Light reallocation for high contrast projection using an analog micromirror array. In *ACM Transactions on Graphics (TOG)* (Vol. 29, No. 6, p. 165). ACM.

Jacquemin, C., Chan, W. K., & Courgeon, M. (2010, October). Bateau ivre: an artistic markerless outdoor mobile augmented reality installation on a riverboat. In *Proceedings of the international conference on Multimedia* (pp. 1353-1362). ACM.

Joffroy, C., Caramel, B., Dery-Pinna, A. M., & Riveill, M. (2011, June). When the functional composition drives the user interfaces composition: process and formalization. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 207-216). ACM.

Karhulahti, V. M. (2012, September). Double fine adventure and the double hermeneutic videogame. In *Proceedings of the 4th International Conference on Fun and Games* (pp. 19-26). ACM.

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012, September). Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems* (pp. 43-50). ACM.

Konyha, Z., Lež, A., Matković, K., Jelović, M., & Hauser, H. (2012, September). Interactive visual analysis of families of curves using data aggregation and derivation. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (p. 24). ACM.

Kristensson, P. O., & Denby, L. C. (2011, August). Continuous recognition and visualization of pen strokes and touch-screen gestures. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling* (pp. 95-102). ACM.

Kulik, A., Kunert, A., Beck, S., Reichel, R., Blach, R., Zink, A., & Froehlich, B. (2011, December). C1x6: a stereoscopic six-user display for co-located collaboration in shared virtual environments. In *ACM Transactions on Graphics (TOG)* (Vol. 30, No. 6, p. 188). ACM.

Lai, C. H., Niinimäki, M., Tahiroglu, K., Kildal, J., & Ahmaniemi, T. (2011, November). Perceived physicality in audio-enhanced force input. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 287-294). ACM.

Martinie, C., Palanque, P., Navarre, D., Winckler, M., & Poupart, E. (2011, June). Model-based training: an approach supporting operability of critical interactive systems. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 53-62). ACM.

Moriya, Y., Tanaka, T., Miyajima, T., & Fujita, K. (2012, August). Estimation of conversational activation level during video chat using turn-taking information. In *Proceedings of the 10th asia pacific conference on Computer human interaction* (pp. 289-298). ACM.

O'Keeffe, I., Staikopoulos, A., Rafter, R., Walsh, E., Yousuf, B., Conlan, O., & Wade, V. (2012, September). Personalized activity based eLearning. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (p. 2). ACM.

Ra, M. R., Sheth, A., Mummert, L., Pillai, P., Wetherall, D., & Govindan, R. (2011, June). Odessa: enabling interactive perception applications on mobile devices. In *Proceedings of the 9th international conference on Mobile systems, applications, and services* (pp. 43-56). ACM.

Rugg, B. M. (2011, November). Charting a new course from blackboard to sakai. In *Proceedings of the 39th ACM annual conference on SIGUCCS* (pp. 53-60). ACM.

Tissoires, B., & Conversy, S. (2011, June). Hayaku: designing and optimizing finely tuned and portable interactive graphics with a graphical compiler. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 117-126). ACM.

Tjin-Kam-Jet, K., Trieschnigg, D., & Hiemstra, D. (2012, August). An analysis of free-text queries for a multi-field web form. In *Proceedings of the 4th Information Interaction in Context Symposium* (pp. 82-89). ACM.

Trattner, C., Helic, D., Singer, P., & Strohmaier, M. (2012, September). Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (p. 14). ACM.

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A. C., Gerber, D., & Cimiano, P. (2012, April). Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web* (pp. 639-648). ACM.

Vaaraniemi, M., Treib, M., & Westermann, R. (2012, July). Temporally coherent real-time labeling of dynamic scenes. In *Proceedings of the 3rd International Conference on Computing for*

Geospatial Research and Applications (p. 17). ACM.

Vennelakanti, R., Dey, P., Shekhawat, A., & Pisupati, P. (2011, November). The picture says it all!: multimodal interactions and interaction metadata. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 89-96). ACM.

Xie, B., Watkins, I., & Huang, M. (2010, November). Older adults' perceptions and use of web-based multimedia health tutorials. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 565-574). ACM.

Information Systems

Afshani, P. (2012, June). Improved pointer machine and I/O lower bounds for simplex range reporting and related problems. In *Proceedings of the 2012 symposuim on Computational Geometry* (pp. 339-346). ACM.

Bidoit, N., Colazzo, D., Malla, N., & Sartiani, C. (2012, August). Partitioning XML documents for iterative queries. In *Proceedings of the 16th International Database Engineering & Applications Sysmposium* (pp. 51-60). ACM.

Costa, J. P., Martins, P., Cecilio, J., & Furtado, P. (2012, August). TEEPA: a timely-aware elastic parallel architecture. In *Proceedings of the 16th International Database Engineering & Applications Sysmposium* (pp. 24-31). ACM.

Durant, K. T., McCray, A. T., & Safran, C. (2010, November). Modeling the temporal evolution of an online cancer forum. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 356-365). ACM.

Friedland, G., Vinyals, O., & Darrell, T. (2010, October). Multimodal location estimation. In *Proceedings of the international conference on Multimedia* (pp. 1245-1252). ACM.

Guillén, R., Jensen, C., & Edelson, S. (2010, November). A machine learning approach for identifying subtypes of autism. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 620-628). ACM.

Hasan, S., O'Riain, S., & Curry, E. (2012, July). Approximate semantic matching of heterogeneous events. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems* (pp. 252-263). ACM.

Hirzel, M. (2012, July). Partition and compose: Parallel complex event processing. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems* (pp. 191-200). ACM.

Iyenghar, P., Westerkamp, C., Wuebbelmann, J., & Pulvermueller, E. (2010, October). A model based approach for debugging embedded systems in real-time. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 69-78). ACM.

Kientz, J. A. (2010, November). Understanding parent-pediatrician interactions for the design of health technologies. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 230-239). ACM.

Larsen, K. G. (2012, May). The cell probe complexity of dynamic range counting. In *Proceedings of the 44th symposium on Theory of Computing* (pp. 85-94). ACM.

Marinos, D., Geiger, C., & Herder, J. (2012, July). Large-area moderator tracking and demonstrational configuration of position based interactions for virtual studios. In *Proceedings of the 10th European conference on Interactive tv and video* (pp. 105-114). ACM.

Matsuo, Y., Shimosawa, T., & Ishikawa, Y. (2012, June). A file I/O system for many-core based clusters. In *Proceedings of the 2nd International Workshop on Runtime and Operating Systems for Supercomputers* (p. 3). ACM.

Mohapatra, A., & Genesereth, M. (2012, August). Incrementally maintaining run-length encoded attributes in column stores. In *Proceedings of the 16th International Database Engineering & Applications Sysmposium* (pp. 146-154). ACM.

Nadimpalli, S. V., & Kumari, V. V. (2012, August). Detecting dependencies in an anonymized dataset. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 82-89). ACM.

Palpanas, T. (2012). A knowledge mining framework for business analysts.*ACM SIGMIS Database*, *43*(1), 46-60.

Plant, C., Sorg, C., Riedl, V., & Wohlschläger, A. (2011, August). Homogeneity-based feature extraction for classification of early-stage alzheimer's disease from functional magnetic resonance images. In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare* (pp. 33-41). ACM.

Silva, Y. N., Reed, J. M., & Tsosie, L. M. (2012, August). MapReduce-based similarity join for metric spaces. In *Proceedings of the 1st International Workshop on Cloud Intelligence* (p. 3). ACM.

Simitsis, A., Wilkinson, K., Castellanos, M., & Dayal, U. (2012, May). Optimizing analytic data flows for multiple execution engines. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 829-840). ACM.

van den Braak, S. W., Choenni, S., Meijer, R., & Zuiderwijk, A. (2012, June). Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (pp. 135-144). ACM.

Vydiswaran, V. G., Zhai, C., & Roth, D. (2011, August). Gauging the internet doctor: ranking medical claims based on community knowledge. In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare* (pp. 42-51). ACM.

Walsh, E., O'Connor, A., & Wade, V. (2012, June). Evaluation of a domain-aware approach to user model interoperability. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 197-206). ACM.

Zhang, M., Zhang, N., & Das, G. (2012, May). Aggregate suppression for enterprise search engines. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 469-480). ACM.

Mathematics of computing

Bonchi, F., Perego, R., Silvestri, F., Vahabi, H., & Venturini, R. (2012, August). Efficient query recommendations in the long tail via center-piece subgraphs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 345-354). ACM.

Brodal, G. S., Lagogiannis, G., & Tarjan, R. E. (2012, May). Strict fibonacci heaps. In *Proceedings of the 44th symposium on Theory of Computing* (pp. 1177-1184). ACM.

Cheng, Y., Xie, Y., Zhang, K., Agrawal, A., & Choudhary, A. (2012, August). CluChunk: clustering large scale user-generated content incorporating chunklet information. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications* (pp. 12-19). ACM.

Colin de Verdière, É., Ginot, G., & Goaoc, X. (2012, June). Multinerves and helly numbers of acyclic families. In *Proceedings of the 2012 Symposium on Computational Geometry* (pp. 209-218). ACM.

Dinitz, M., & Krauthgamer, R. (2011, June). Fault-tolerant spanners: better and simpler. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS Symposium on Principles of distributed computing* (pp. 169-178). ACM.

El-Arini, K., Paquet, U., Herbrich, R., Van Gael, J., & Agüera y Arcas, B. (2012, August). Transparent user models for personalization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 678-686). ACM.

Eliáš, M., & Matoušek, J. (2012). Higher-order Erdos-Szekeres theorems. In *Proceedings of the 28th ACM Symposium on Computational Geometry* (pp. 81-90). ACM

Grossi, R., & Ottaviano, G. (2012, May). The wavelet trie: Maintaining an indexed sequence of strings in compressed space. In *Proceedings of the 31st symposium on Principles of Database Systems* (pp. 203-214). ACM.

Gupta, M., Gao, J., Sun, Y., & Han, J. (2012, August). Integrating community matching and outlier detection for mining evolutionary community outliers. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 859-867). ACM.

Han, L., Song, G., Cong, G., & Xie, K. (2012, August). Overlapping decomposition for causal graphical modeling. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 114-122). ACM.

Khabbazian, M., & Kowalski, D. R. (2011, June). Time-efficient randomized multiple-message broadcast in radio networks. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 373-380). ACM.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012, August). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 786-794). ACM.

Korman, A., Kutten, S., & Masuzawa, T. (2011, June). Fast and compact self stabilizing verification, computation, and fault detection of an MST. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 311-320). ACM.

Korman, A., Sereni, J. S., & Viennot, L. (2011, June). Toward more localized local algorithms: removing assumptions concerning global knowledge. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 49-58). ACM.

Kothapalli, K., & Pemmaraju, S. (2011, June). Distributed graph coloring in a few rounds. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 31-40). ACM.

Lenzen, C., & Wattenhofer, R. (2011, June). MIS on Trees. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 41-48). ACM.

Lejsek, H., Jónsson, B. Þ., & Amsaleg, L. (2011, April). NV-Tree: nearest neighbors at the billion scale. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (p. 54). ACM.

Martens, C., & Crary, K. (2012, September). LF in LF: mechanizing the metatheories of LF in twelf. In *Proceedings of the seventh international workshop on Logical frameworks and meta-languages, theory and practice* (pp. 23-32). ACM.

Nakatani, M., Ohno, T., Nakane, A., Komatsubara, A., & Hashimoto, S. (2012, August). How to motivate people to use internet at home: understanding the psychology of non-active users. In *Proceedings of the 10th asia pacific conference on Computer human interaction* (pp. 259-268). ACM.

Pettarin, A., Pietracaprina, A., Pucci, G., & Upfal, E. (2011, June). Tight bounds on information dissemination in sparse mobile networks. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 355-362). ACM.

Prokopec, A., Bronson, N. G., Bagwell, P., & Odersky, M. (2012, February). Concurrent tries with efficient non-blocking snapshots. In *ACM SIGPLAN Notices* (Vol. 47, No. 8, pp. 151-160). ACM.

Sato, I., Kurihara, K., & Nakagawa, H. (2012, August). Practical collapsed variational bayes inference for hierarchical dirichlet process. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 105-113). ACM.

Sitchinava, N., & Zeh, N. (2012, June). A parallel buffer tree. In *Proceedings of the 24th ACM symposium on Parallelism in algorithms and architectures* (pp. 214-223). ACM.

Whitley, D., & Chen, W. (2012, July). Constant time steepest descent local search with lookahead for NK-landscapes and MAX-kSAT. In *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference* (pp. 1357-1364). ACM.

Networks

Chen, Y., Borrel, V., Ammar, M., & Zegura, E. (2011). A framework for characterizing the wireless

and mobile network continuum. *ACM SIGCOMM Computer Communication Review*, *41*(1), 5-13.

Chauhan, J., & Makaroff, D. (2012, June). Performance evaluation of video-on-demand in virtualized environments: the client perspective. In *Proceedings of the 6th international workshop on Virtualization Technologies in Distributed Computing Date* (pp. 29-36). ACM.

Choochaisri, S., Apicharttrisorn, K., Korprasertthaworn, K., Taechalertpaisarn, P., & Intanagonwiwat, C. (2012). Desynchronization with an artificial force field for wireless networks. *ACM SIGCOMM Computer Communication Review*, *42*(2), 7-15.

Dunn, C. W., Gupta, M., Gerber, A., & Spatscheck, O. (2012, August). Navigation characteristics of online social networks and search engines users. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks* (pp. 43-48). ACM.

Fusco, F., Dimitropoulos, X., Vlachos, M., & Deri, L. (2012). pcapIndex: an index for network packet traces with legacy compatibility. *ACM SIGCOMM Computer Communication Review*, *42*(1), 47-53.

Garay, J. A., Katz, J., Kumaresan, R., & Zhou, H. S. (2011, June). Adaptively secure broadcast, revisited. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 179-186). ACM.

Ghodsi, A., Sekar, V., Zaharia, M., & Stoica, I. (2012). Multi-resource fair queueing for packet processing. *ACM SIGCOMM Computer Communication Review*, *42*(4), 1-12.

Huang, C., Batanov, I., & Li, J. (2012). A practical solution to the client-LDNS mismatch problem. *ACM SIGCOMM Computer Communication Review*, *42*(2), 35-41.

Koponen, T., Shenker, S., Balakrishnan, H., Feamster, N., Ganichev, I., Ghodsi, A., ... & Kuptsov, D. (2011). Architecting for innovation. *ACM SIGCOMM Computer Communication Review*, *41*(3), 24-36.

Küsters, R., & Tuengerthal, M. (2011, October). Composition theorems without pre-established session identifiers. In *Proceedings of the 18th ACM conference on Computer and communications security* (pp. 41-50). ACM.

Lin, K., Chuang, Y. J., & Katabi, D. (2012). A light-weight wireless handshake. *ACM SIGCOMM Computer Communication Review*, *42*(2), 28-34.

Ma, Y., & Banerjee, S. (2012, August). A smart pre-classifier to reduce power consumption of TCAMs for multi-dimensional packet classification. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication* (pp. 335-346). ACM.

Marzolla, M. (2012, June). Optimizing the energy consumption of large-scale applications. In *Proceedings of the 8th international ACM SIGSOFT conference on Quality of Software Architectures* (pp. 123-132). ACM.

Miller, K., Sanne, A., Srinivasan, K., & Vishwanath, S. (2012, June). Enabling real-time interference alignment: promises and challenges. In *Proceedings of the thirteenth ACM*

international symposium on Mobile Ad Hoc Networking and Computing (pp. 55-64). ACM.

Mittal, R., Kansal, A., & Chandra, R. (2012, August). Empowering developers to estimate app energy consumption. In *Proceedings of the 18th annual international conference on Mobile computing and networking* (pp. 317-328). ACM.

Pandurangan, G., & Trehan, A. (2011, June). Xheal: localized self-healing using expanders. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 301-310). ACM.

Rane, A., Browne, J., & Koesterke, L. (2012, July). A systematic process for efficient execution on Intel's heterogeneous computation nodes. In *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond* (p. 8). ACM.

Rétvári, G., Gulyás, A., Heszberger, Z., Csernai, M., & Bíró, J. J. (2011, June). Compact policy routing. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 149-158). ACM.

Sarrar, N., Uhlig, S., Feldmann, A., Sherwood, R., & Huang, X. (2012). Leveraging Zipf's law for traffic offloading. *ACM SIGCOMM Computer Communication Review*, *42*(1), 16-22.

Shue, C. A., Kalafut, A. J., Allman, M., & Taylor, C. R. (2012). On building inexpensive network capabilities. *ACM SIGCOMM Computer Communication Review*, *42*(2), 72-79.

Vesco, A., Abrate, F., & Scopigno, R. (2011, October). Convergence and performance analysis of leaderless synchronization in Wi-Fi networks. In *Proceedings of the 6th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks* (pp. 49-58). ACM.

Wang, J., Hassanieh, H., Katabi, D., & Indyk, P. (2012). Efficient and reliable low-power backscatter networks. *ACM SIGCOMM Computer Communication Review*, *42*(4), 61-72.

Wang, R., Maciocco, C., Tai, T. Y., Yavatkar, R., Lu, L. K., & Min, A. W. (2012, May). DirectPath: high performance and energy efficient platform I/O architecture for content intensive usages. In *Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy), 2012 Third International Conference on* (pp. 1-10). IEEE.

Westermann, D., Happe, J., Krebs, R., & Farahbod, R. (2012, September). Automated inference of goal-oriented performance prediction functions. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering* (pp. 190-199). ACM.

Yin, D., Unnikrishnan, D., Liao, Y., Gao, L., & Tessier, R. (2011). Customizing virtual networks with partial FPGA reconfiguration. *ACM SIGCOMM Computer Communication Review*, *41*(1), 125-132.

Zave, P. (2012). Using lightweight modeling to understand chord. *ACM SIGCOMM Computer Communication Review*, *42*(2), 49-57.

Zhu, J., & Hajek, B. (2012). Stability of a peer-to-peer communication system. *Information Theory,*

IEEE Transactions on, *58*(7), 4693-4713.

Security and privacy

Adams, E. K., Intwala, M., & Kapadia, A. (2010, November). MeD-Lights: a usable metaphor for patient controlled access to electronic health records. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 800-808). ACM.

AL Faresi, A., Wijesekera, D., & Moidu, K. (2010, November). A comprehensive privacy-aware authorization framework founded on HIPAA privacy rules. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 637-646). ACM.

Ali, S. T., Sivaraman, V., & Ostry, D. (2012, April). Zero reconciliation secret key generation for body-worn health monitoring devices. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks* (pp. 39-50). ACM.

Chen, F., Liu, A. X., Hwang, J., & Xie, T. (2012). First step towards automatic correction of firewall policy faults. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, *7*(2), 27.

Clement, A., Junqueira, F., Kate, A., & Rodrigues, R. (2012, July). On the (limited) power of non-equivocation. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing* (pp. 301-308). ACM.

Erkin, Z., Beye, M., Veugen, T., & Lagendijk, R. L. (2012, September). Privacy-preserving content-based recommender system. In *Proceedings of the on Multimedia and security* (pp. 77-84). ACM.

Fan, L., Wang, Y., Cheng, X., & Jin, S. (2012, August). Quantitative analysis for privacy leak software with privacy Petri net. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics* (p. 7). ACM.

Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J., & Grandison, T. (2010, November). An evaluation of feature sets and sampling techniques for de-identification of medical records. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 183-190). ACM.

Grace, M. C., Zhou, W., Jiang, X., & Sadeghi, A. R. (2012, April). Unsafe exposure analysis of mobile in-app advertisements. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks* (pp. 101-112). ACM.

Grace, M., Zhou, Y., Zhang, Q., Zou, S., & Jiang, X. (2012, June). Riskranker: scalable and accurate zero-day android malware detection. In *Proceedings of the 10th international conference on Mobile systems, applications, and services* (pp. 281-294). ACM.

Groth, J., Ostrovsky, R., & Sahai, A. (2012). New techniques for noninteractive zero-knowledge. *Journal of the ACM (JACM)*, *59*(3), 11.

Hoens, T. R., Blanton, M., & Chawla, N. V. (2010, November). Reliable medical recommendation systems with patient privacy. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 173-182). ACM.

Löhr, H., Sadeghi, A. R., & Winandy, M. (2010, November). Securing the e-health cloud. In

Proceedings of the 1st ACM International Health Informatics Symposium (pp. 220-229). ACM.

López-Alt, A., Tromer, E., & Vaikuntanathan, V. (2012, May). On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In *Proceedings of the 44th symposium on Theory of Computing* (pp. 1219-1234). ACM.

Martínez-Prieto, M. A., Fernández, J. D., & Cánovas, R. (2012). Querying RDF dictionaries in compressed space. *ACM SIGAPP Applied Computing Review*, *12*(2), 64-77.

Martino, L., & Ahuja, S. (2010, November). Privacy policies of personal health records: an evaluation of their effectiveness in protecting patient information. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 191-200). ACM.

Naehrig, M., Lauter, K., & Vaikuntanathan, V. (2011, October). Can homomorphic encryption be practical?. In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop* (pp. 113-124). ACM.

Srinivasan, M. K., Sarukesi, K., Rodrigues, P., Manoj, M. S., & Revathy, P. (2012, August). State-of-the-art cloud computing security taxonomies: a classification of security challenges in the present cloud computing environment. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 470-476). ACM.

Teixeira, A., Pérez, D., Sandberg, H., & Johansson, K. H. (2012, April). Attack models and scenarios for networked control systems. In *Proceedings of the 1st international conference on High Confidence Networked Systems* (pp. 55-64). ACM.

Wei, S., & Potkonjak, M. (2012, April). Wireless security techniques for coordinated manufacturing and on-line hardware trojan detection. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks* (pp. 161-172). ACM.

Wright, N., Patrick, A. S., & Biddle, R. (2012, July). Do you see your password?: applying recognition to textual passwords. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (p. 8). ACM.

Zeng, Y., Shin, K. G., & Hu, X. (2012, April). Design of SMS commanded-and-controlled and P2P-structured mobile botnets. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks* (pp. 137-148). ACM.

Zhu, Y., Hu, H., Ahn, G. J., Yu, M., & Zhao, H. (2012, February). Comparison-based encryption for fine-grained access control in clouds. In *Proceedings of the second ACM conference on Data and Application Security and Privacy* (pp. 105-116). ACM.

Software and its engineering

Althaus, E., Altmeyer, S., & Naujoks, R. (2011, April). Precise and efficient parametric path analysis. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 141-150). ACM.

Anta, A., Majumdar, R., Saha, I., & Tabuada, P. (2010, October). Automatic verification of control system implementations. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 9-18). ACM.

Avrunin, G. S., Clarke, L. A., Osterweil, L. J., Christov, S. C., Chen, B., Henneman, E. A., ... & Mertens, W. (2010, November). Experience modeling and analyzing medical processes: UMass/baystate medical safety project overview. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 316-325). ACM.

Barbosa, A., Paiva, A. C., & Campos, J. C. (2011, June). Test case generation from mutated task models. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 175-184). ACM.

Buckl, C., Gaponova, I., Geisinger, M., Knoll, A., & Lee, E. A. (2010, October). Model-based specification of timing requirements. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 239-248). ACM.

Chang, L. P., & Huang, L. C. (2011, April). A low-cost wear-leveling algorithm for block-mapping solid-state disks. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 31-40). ACM.

Chang, Y. H., & Kuo, T. W. (2010, October). A reliable MTD design for MLC flash-memory storage systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 179-188). ACM.

Craciunas, S. S., Kirsch, C. M., & Sokolova, A. (2010, October). Power-aware temporal isolation with variable-bandwidth servers. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 259-268). ACM.

Dobre, D., Guerraoui, R., Majuntke, M., Suri, N., & Vukolić, M. (2011, June). The complexity of robust atomic storage. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 59-68). ACM.

Falerio, J. M., Rajamani, S., Rajan, K., Ramalingam, G., & Vaswani, K. (2012, July). Generalized lattice agreement. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing* (pp. 125-134). ACM.

Fung, K. H., & Low, G. C. (2011). Quality factors for dynamic evolution in composition-based distributed applications. *ACM SIGMIS Database*, *42*(1), 29-58.

Gamatié, A., & Gonnord, L. (2011). Static analysis of synchronous programs in signal for efficient design of multi-clocked embedded systems. *ACM SIGPLAN Notices*, *46*(5), 71-80.

Henry, G., Mauny, M., Chailloux, E., & Manoury, P. (2012, September). Typing unmarshalling without marshalling types. In *Proceedings of the 17th ACM SIGPLAN international conference on Functional programming* (pp. 287-298). ACM.

Iyenghar, P., Westerkamp, C., Wuebbelmann, J., & Pulvermueller, E. (2010, October). A model based approach for debugging embedded systems in real-time. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 69-78). ACM.

Kähkönen, K., Saarikivi, O., & Heljanko, K. (2012, September). Using unfoldings in automated testing of multithreaded programs. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering* (pp. 150-159). ACM.

Kwon, H., Kim, E., Choi, J., Lee, D., & Noh, S. H. (2010, October). Janus-FTL: finding the optimal point on the spectrum between page and block mapping schemes. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 169-178). ACM.

Li, H., & Baruah, S. (2010, October). Load-based schedulability analysis of certifiable mixed-criticality systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 99-108). ACM.

Moscibroda, T., & Oshman, R. (2011, June). Resilience of mutual exclusion algorithms to transient memory faults. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 69-78). ACM.

Myreen, M. O., & Owens, S. (2012, September). Proof-producing synthesis of ML from higher-order logic. In *Proceedings of the 17th ACM SIGPLAN international conference on Functional programming* (pp. 115-126). ACM.

Sarkar, A., Mueller, F., & Ramaprasad, H. (2011, April). Predictable task migration for locked caches in multi-core systems. In *ACM SIGPLAN Notices* (Vol. 46, No. 5, pp. 131-140). ACM.

Thomas, J. J., Fischmeister, S., & Kumar, D. (2011). Lowering overhead in sampling-based execution monitoring and tracing. *ACM SIGPLAN Notices*, *46*(5), 101-110.

Van den Bergh, J., Luyten, K., & Coninx, K. (2011, June). CAP3: context-sensitive abstract user interface specification. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 31-40). ACM.

Wahba, S. K., Hallstrom, J. O., & Soundarajan, N. (2010, October). Initiating a design pattern catalog for embedded network systems. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 249-258). ACM.

Theory of computation

Almorsy, M., Grundy, J., & Ibrahim, A. S. (2012, September). Supporting automated vulnerability analysis using formalized vulnerability signatures. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering* (pp. 100-109). ACM.

Avrunin, G. S., Clarke, L. A., Osterweil, L. J., Christov, S. C., Chen, B., Henneman, E. A., ... & Mertens, W. (2010, November). Experience modeling and analyzing medical processes: UMass/baystate medical safety project overview. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 316-325). ACM.

Bansal, N., Lee, K. W., Nagarajan, V., & Zafer, M. (2011, June). Minimum congestion mapping in a cloud. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 267-276). ACM.

Benveniste, A., Bouillard, A., & Caspi, P. (2010, October). A unifying view of loosely time-triggered architectures. In *Proceedings of the tenth ACM international conference on Embedded software* (pp. 189-198). ACM.

Bhalgat, A., Feldman, J., & Mirrokni, V. (2012, August). Online allocation of display ads with smooth delivery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1213-1221). ACM.

Censor-Hillel, K., Gilbert, S., Kuhn, F., Lynch, N., & Newport, C. (2011, June). Structuring unreliable radio networks. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 79-88). ACM.

Cervesato, I., Pfenning, F., Sacchini, J. L., Schürmann, C., & Simmons, R. J. (2012, September). Trace matching in a concurrent logical framework. In *Proceedings of the seventh international workshop on Logical frameworks and meta-languages, theory and practice* (pp. 1-12). ACM.

Chen, B., Yu, H., Zhao, Y., & Gibbons, P. B. (2012, July). The cost of fault tolerance in multi-party communication complexity. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing* (pp. 57-66). ACM.

Dinh, T. N., Nguyen, D. T., & Thai, M. T. (2012, June). Cheap, easy, and massively effective viral marketing in social networks: truth or fiction?. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 165-174). ACM.

Helmi, M., Higham, L., Pacheco, E., & Woelfel, P. (2011, June). The space complexity of long-lived and one-shot timestamp implementations. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 139-148). ACM.

Holzer, S., Locher, T., Pignolet, Y. A., & Wattenhofer, R. (2012, June). Deterministic multi-channel information exchange. In *Proceedings of the 24th ACM symposium on Parallelism in algorithms and architectures* (pp. 109-120). ACM.

Jaffe, A., Moscibroda, T., Effinger-Dean, L., Ceze, L., & Strauss, K. (2011, June). The impact of memory models on software reliability in multiprocessors. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 89-98). ACM.

Krishnaswami, N. R., Turon, A., Dreyer, D., & Garg, D. (2012, September). Superficially substructural types. In *Proceedings of the 17th ACM SIGPLAN international conference on Functional programming* (pp. 41-54). ACM.

Kuhn, F., Oshman, R., & Moses, Y. (2011, June). Coordinated consensus in dynamic networks. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 1-10). ACM.

Liang, G., & Vaidya, N. (2011, June). Error-free multi-valued consensus with byzantine failures. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 11-20). ACM.

Sacco, O., & Breslin, J. G. (2012, September). PPO & PPM 2.0: Extending the Privacy Preference Framework to provide finer-grained access control for the Web of Data. In *Proceedings of the 8th International Conference on Semantic Systems* (pp. 80-87). ACM.

Severi, P. G., & de Vries, F. J. J. (2012, September). Pure type systems with corecursion on streams: from finite to infinitary normalisation. In *Proceedings of the 17th ACM SIGPLAN international*

conference on Functional programming (pp. 141-152). ACM.

Tao, Y. (2012, May). Indexability of 2D range search revisited: constant redundancy and weak indivisibility. In *Proceedings of the 31st symposium on Principles of Database Systems* (pp. 131-142). ACM.

Wadler, P. (2012, September). Propositions as sessions. In *Proceedings of the 17th ACM SIGPLAN international conference on Functional programming* (pp. 273-286). ACM.

Appendix D: CSJAC Bibliography

Computer systems organisation

Bugnion, E., Devine, S., Rosenblum, M., Sugerman, J., & Wang, E. Y. (2012). Bringing Virtualization to the x86 Architecture with the Original VMware Workstation. *ACM Transactions on Computer Systems (TOCS)*, 30(4), 12.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2), 4.

Di Biagio, A., Agosta, G., Sykora, M., & Silvano, C. (2012). Architecture Optimization of Application-Specific Implicit Instructions. *ACM Transactions on Embedded Computing Systems (TECS)*, 11(S2), 44.

Dobson, S., Denazis, S., Fernández, A., Gaïti, D., Gelenbe, E., Massacci, F., ... & Zambonelli, F. (2006). A survey of autonomic communications. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 1(2), 223-259.

Fowers, J., Brown, G., Wernsing, J., & Stitt, G. (2013). A performance and energy comparison of convolution on GPUs, FPGAs, and multicore processors. *ACM Transactions on Architecture and Code Optimization (TACO)*, 9(4), 25.

Galuzzi, C., & Bertels, K. (2008). The instruction-set extension problem: A survey. In *Reconfigurable Computing: Architectures, Tools and Applications*(pp. 209-220). Springer Berlin Heidelberg.

Govindan, S., Wang, D., Sivasubramaniam, A., & Urgaonkar, B. (2013). Aggressive Datacenter Power Provisioning with Batteries. *ACM Transactions on Computer Systems (TOCS)*, 31(1), 2.

Han, K., Ahn, J., & Choi, K. (2013). Power-Efficient Predication Techniques for Acceleration of Control Flow Execution on CGRA. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(2), 8.

Jackson, B. L., Rajendran, B., Corrado, G. S., Breitwisch, M., Burr, G. W., Cheek, R., ... & Modha, D. S. (2013). Nanoscale electronic synapses using phase change devices. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 9(2), 12.

Jayaram, K. R., Eugster, P., & Jayalath, C. (2013). Parametric Content-Based Publish/Subscribe. *ACM Transactions on Computer Systems (TOCS)*, 31(2), 4.

Jiang, H., & Hallstrom, J. O. (2011). Fast, accurate event classification on resource-lean embedded sensors. In *Wireless Sensor Networks* (pp. 65-80). Springer Berlin Heidelberg.

Kritikakou, A., Catthoor, F., Athanasiou, G. S., Kelefouras, V., & Goutis, C. (2013). Near-Optimal Microprocessor and Accelerators Codesign with Latency and Throughput Constraints. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(2), 6.

Li, S., Ahn, J. H., Strong, R. D., Brockman, J. B., Tullsen, D. M., & Jouppi, N. P. (2013). The

McPAT Framework for Multicore and Manycore Architectures: Simultaneously Modeling Power, Area, and Timing. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(1), 5.

Schneider, D., & Trapp, M. (2013). Conditional Safety Certification of Open Adaptive Systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(2), 8.

Schuhmann, S., Herrmann, K., Rothermel, K., & Boshmaf, Y. (2013). Adaptive Composition of Distributed Pervasive Applications in Heterogeneous Environments. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)–to appear.*

Sheikh, H. F., Tan, H., Ahmad, I., Ranka, S., & Bv, P. (2012). Energy-and performance-aware scheduling of tasks on parallel and distributed systems.*ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 8(4), 32.

Yeh, C. C., Chang, K. C., Chen, T. F., & Yeh, C. (2011). Maintaining performance on power gating of microprocessor functional units by using a predictive pre-wakeup strategy. *ACM Transactions on Architecture and Code Optimization (TACO)*, 8(3), 16.

Computing methodologies

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Črepinšek, M., Liu, S. H., & Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: a survey. *ACM Computing Surveys (CSUR)*, 45(3), 35.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264-323.

Joglekar, N. R., Anderson, E. G., & Shankaranarayanan, G. (2013). Accuracy of aggregate data in distributed project settings: Model, analysis and implications. *Journal of Data and Information Quality (JDIQ)*, 4(3), 13.

Gallacher, S., Papadopoulou, E., Taylor, N. K., & Williams, M. H. (2013). Learning user preferences for adaptive pervasive environments: An incremental and temporal approach. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(1), 5.

Hoseini-Tabatabaei, S. A., Gluhak, A., & Tafazolli, R. (2013). A survey on smartphone-based systems for opportunistic user context recognition. *ACM Computing Surveys (CSUR)*, 45(3), 27.

Rotta, R., & Noack, A. (2011). Multilevel local search algorithms for modularity clustering. *Journal of Experimental Algorithmics (JEA)*, 16, 2-3.

Serban, F., Vanschoren, J., Kietz, J. U., & Bernstein, A. (2012). A survey of intelligent assistants for data analysis. *ACM Computing Surveys*.

Song, W., Finch, A., Tanaka-Ishii, K., Yasuda, K., & Sumita, E. (2013). picoTrans: An intelligent icon-driven interface for cross-lingual communication. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1), 5.

Song, Y., Demirdjian, D., & Davis, R. (2012). Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1), 5.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), 13.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4), 399-458.

Hardware

Apalkov, D., Khvalkovskiy, A., Watts, S., Nikitin, V., Tang, X., Lottis, D., ... & Krounbi, M. (2013). Spin-transfer torque magnetic random access memory (STT-MRAM). *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 9(2), 13.

Beausoleil, R. G. (2011). Large-scale integrated photonics for high-performance interconnects. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 7(2), 6.

Chatterjee, S., Salahuddin, S., Kumar, S., & Mukhopadhyay, S. (2013). Electrothermal analysis of spin-transfer-torque random access memory arrays. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *9*(2), 15.

Chen, Y., Wong, W. F., Li, H., Koh, C. K., Zhang, Y., & Wen, W. (2013). On-chip caches built on multilevel spin-transfer torque RAM cells and its optimizations. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *9*(2), 16.

Dehon, A. (2005). Nanowire-based programmable architectures. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 1(2), 109-162.

Dong, W., & Li, P. (2011). Parallel circuit simulation with adaptively controlled projective integration. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 16(4), 44.

Jiang, L., Du, Y., Zhao, B., Zhang, Y., Childers, B. R., & Yang, J. (2013). Hardware-Assisted Cooperative Integration of Wear-Leveling and Salvaging for Phase Change Memory. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(2), 7.

Lee, S. W., Park, D. J., Chung, T. S., Lee, D. H., Park, S., & Song, H. J. (2007). A log buffer-based flash translation layer using fully-associative sector translation. *ACM Transactions on Embedded Computing Systems (TECS)*, 6(3), 18.

Mojumder, N. N., Fong, X., Augustine, C., Gupta, S. K., Choday, S. H., & Roy, K. (2013). Dual pillar spin-transfer torque MRAMs for low power applications. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 9(2), 14.

Qian, H., Sapatnekar, S. S., & Kursun, E. (2012). Fast Poisson solvers for thermal analysis. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 17(3), 32.

Qu, G., & Potkonjak, M. (2003). System synthesis of synchronous multimedia applications. *ACM Transactions on Embedded Computing Systems (TECS)*, 2(1), 74-97.

Roy, S., Mitra, D., Bhattacharya, B. B., & Chakrabarty, K. (2012). Congestion-aware layout design for high-throughput digital microfluidic biochips. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *8*(3), 17.

Suresh, D. C., Agrawal, B., Yang, J., & Najjar, W. (2009). Energy-efficient encoding techniques for off-chip data buses. *ACM Transactions on Embedded Computing Systems (TECS)*, 8(2), 9.

Udayakumaran, S., Dominguez, A., & Barua, R. (2006). Dynamic allocation for scratch-pad memory using compile-time decisions. *ACM Transactions on Embedded Computing Systems (TECS)*, *5*(2), 472-511.

Voros, N. S., Hübner, M., Becker, J., Kühnle, M., Thomaitiv, F., Grasset, A., ... & Putzke-Röming, W. (2013). MORPHEUS: A heterogeneous dynamically reconfigurable platform for designing highly complex embedded systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 12(3), 70.

Yang, J. J., & Williams, R. S. (2013). Memristive devices in computing system: Promises and challenges. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 9(2), 11.

Zang, W., & Gordon-Ross, A. (2013). A survey on cache tuning from a power/energy perspective. *ACM Computing Surveys (CSUR)*, 45(3), 32.

Human-centred computing

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293-327.

Biran, D., Zack, M. H., & Briotta, R. J. (2013). Competitive intelligence and information quality: A game-theoretic perspective. *Journal of Data and Information Quality (JDIQ)*, 4(3), 12.

Carroll, J. M., & Rosson, M. B. (2013). Wild at Home: The Neighborhood as a Living Laboratory for HCI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(3), 16.

Console, L., Antonelli, F., Biamino, G., Carmagnola, F., Cena, F., Chiabrando, E., ... & Vernero, F. (2013). Interacting with social networks of intelligent things and people in the world of gastronomy. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1), 4.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3), 18.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects?. *ACM Transactions on Graphics (TOG)*, 31(4), 44.

Gingold, Y., Shamir, A., & Cohen-Or, D. (2012). Micro perceptual human computation for visual tasks. *ACM Transactions on Graphics (TOG)*, 31(5), 119.

Hassenzahl, M., Heidecker, S., Eckoldt, K., Diefenbach, S., & Hillmann, U. (2012). All You Need is Love: Current Strategies of Mediating Intimate Relationships through Technology. *ACM*

Transactions on Computer-Human Interaction (TOCHI),19(4), 30.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174-196.

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(1), 3.

Malzbender, T., Samadani, R., Scher, S., Crume, A., Dunn, D., & Davis, J. (2012). Printing reflectance functions. *ACM Transactions on Graphics (TOG)*, 31(3), 20.

Nussbaumer, P., Matter, I., & Schwabe, G. (2012). "Enforced" vs."Casual" Transparency--Findings from IT-Supported Financial Advisory Encounters. *ACM Transactions on Management Information Systems (TMIS)*, 3(2), 11.

O'hara, K., Harper, R., Mentis, H., Sellen, A., & Taylor, A. (2013). On the naturalness of touchless: putting the "interaction" back into NUI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *20*(1), 5.

Pradhan, G. N., & Prabhakaran, B. (2012). Analyzing and Visualizing Jump Performance Using Wireless Body Sensors. *ACM Transactions on Embedded Computing Systems (TECS)*, 11(S2), 47.

Rooksby, J. (2013). Wild in the Laboratory: A Discussion of Plans and Situated Actions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(3), 19.

Svanæs, D. (2013). Interaction design for and with the lived body: Some implications of merleau-ponty's phenomenology. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *20*(1), 8.

Yazdani, A., Lee, J. S., Vesin, J. M., & Ebrahimi, T. (2012). Affect recognition based on physiological changes during the watching of music videos. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1), 7.

Information systems

Bellotti, F., Berta, R., De Gloria, A., D'ursi, A., & Fiore, V. (2012). A serious game model for cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(4), 17.

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 22.

Cao, L., & Zhu, H. (2013). Normal accidents: Data quality problems in ERP-enabled manufacturing. *Journal of Data and Information Quality (JDIQ)*, 4(3), 11.

Chan, T. M. (2013). Persistent predecessor search and orthogonal point location on the word RAM. *ACM Transactions on Algorithms (TALG)*, 9(3), 22.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), 15.

Colombo, C. H. R. I. S. T. I. A. N., & Pace, G. (2012). Recovery within long running transactions. *ACM Computing Surveys*, *45*.

Cremonesi, P., Garzotto, F., & Turrin, R. (2012). Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: an Empirical Study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2), 11.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, *40*(2), 5.

Fan, X., Wang, J., Pu, X., Zhou, L., & Lv, B. (2011). On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)*, 2(2), 10.

Harter, T., Dragga, C., Vaughn, M., Arpaci-Dusseau, A. C., & Arpaci-Dusseau, R. H. (2012). A file is not a file: understanding the I/O behavior of Apple desktop applications. *ACM Transactions on Computer Systems (TOCS)*, *30*(3), 10.

Koller, D., Frischer, B., & Humphreys, G. (2009). Research challenges for digital archives of 3D cultural heritage models. *Journal on computing and cultural heritage (JOCCH)*, 2(3), 7.

Lau, R. Y., Liao, S. Y., Kwok, R. C. W., Xu, K., Xia, Y., & Li, Y. (2011). Text mining and probabilistic language modeling for online review spam detecting. *ACM Transactions on Management Information Systems*, 2(4), 1-30.

Müller, H., Freytag, J. C., & Leser, U. (2012). Improving data quality by source analysis. *Journal of Data and Information Quality (JDIQ)*, 2(4), 15.

Mathematics of computing

Batz, G. V., Geisberger, R., Sanders, P., & Vetter, C. (2013). Minimum time-dependent travel times with contraction hierarchies. *Journal of Experimental Algorithmics (JEA)*, 18(1), 1-4.

Bauer, R., & Delling, D. (2009). SHARC: Fast and robust unidirectional routing. *Journal of Experimental Algorithmics (JEA)*, 14, 4.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM (JACM)*, *58*(3), 11.

Damaggio, E., Deutsch, A., & Vianu, V. (2012). Artifact systems with data dependencies and arithmetic. *ACM Transactions on Database Systems (TODS)*, 37(3), 22.

Fussl, A., Fruehwirth-Schnatter, S., & Fruehwirth, R. (2013). Efficient MCMC for Binomial Logit Models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(1), 3.

Görke, R., Maillard, P., Schumm, A., Staudt, C., & Wagner, D. (2013). Dynamic graph clustering combining modularity and smoothness. *Journal of Experimental Algorithmics (JEA)*, 18(1), 1-5.

Kawarabayashi, K. I., & Kobayashi, Y. (2013). An O (log n)-Approximation Algorithm for the Edge-Disjoint Paths Problem in Eulerian Planar Graphs. *ACM Transactions on Algorithms (TALG)*, 9(2), 16.

Łącki, J. (2013). Improved Deterministic Algorithms for Decremental Reachability and Strongly Connected Components. *ACM Transactions on Algorithms (TALG)*, 9(3), 27.

Le Corff, S., & Fort, G. (2013). Convergence of a particle-based approximation of the Block Online Expectation Maximization algorithm. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(1), 2.

Lim, E. (2012). Stochastic approximation over multidimensional discrete sets with applications to inventory systems and admission control of queueing networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *22*(4), 19.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 3-30.

Möhring, R. H., Schilling, H., Schütz, B., Wagner, D., & Willhalm, T. (2007). Partitioning graphs to speedup Dijkstra's algorithm. *Journal of Experimental Algorithmics (JEA)*, 11, 2-8.

Ng, S. H., & Yin, J. (2012). Bayesian kriging analysis and design for stochastic simulations. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *22*(3), 17.

Petkov, V., Rajagopal, R., & Obraczka, K. (2013). Characterizing per-application network traffic using entropy. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(2), 14.

Schreck, A., Fort, G., & Moulines, E. (2013). Adaptive Equi-Energy Sampler: Convergence and Illustration. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(1), 5.

Singh, S. S., Chopin, N., & Whiteley, N. (2013). Bayesian Learning of Noisy Markov Decision Processes. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(1), 4.

Wang, K., Zong, C., & Su, K. Y. (2012). Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2), 7.

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All---A Contingency Approach to Data Governance. *Journal of Data and Information Quality (JDIQ)*, 1(1), 4.

Weimann, O., & Yuster, R. (2013). Replacement Paths and Distance Sensitivity Oracles via Fast Matrix Multiplication. *ACM Transactions on Algorithms (TALG)*, 9(2), 14.

Networks

Ardagna, C. A., Jajodia, S., Samarati, P., & Stavrou, A. (2013). Providing users' anonymity in mobile hybrid networks. *ACM Transactions on Internet Technology (TOIT)*, *12*(3), 7.

Cittadini, L., Battista, G. D., & Rimondini, M. (2012). On the stability of interdomain routing. *ACM Computing Surveys (CSUR)*, 44(4), 26.

Goyal, N., Olver, N., & Shepherd, F. B. (2008, May). The VPN conjecture is true. In *Proceedings of*

the 40th annual ACM symposium on Theory of computing (pp. 443-450). ACM.

Iwanicki, K., & Van Steen, M. (2012). A case for hierarchical routing in low-power wireless embedded networks. *ACM Transactions on Sensor Networks (TOSN)*, 8(3), 25.

Jain, M., & Dovrolis, C. (2002, August). End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. In *ACM SIGCOMM Computer Communication Review* (Vol. 32, No. 4, pp. 295-308). ACM.

Jiang, L., & Walrand, J. (2010). A distributed CSMA algorithm for throughput and utility maximization in wireless networks. *IEEE/ACM Transactions on Networking (TON)*, 18(3), 960-972.

Keys, K., Hyun, Y., Luckie, M., & Claffy, K. (2013). Internet-Scale IPv4 Alias Resolution with MIDAR.

Kim, D., Yoo, S., & Lee, S. (2013). A network congestion-aware memory subsystem for manycore. *ACM Transactions on Embedded Computing Systems (TECS)*, 12(4), 110.

Koksal, C. E., Ercetin, O., & Sarikaya, Y. (2010, November). Control of wireless networks with secrecy. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on* (pp. 47-51). IEEE.

Low, S. H., & Lapsley, D. E. (1999). Optimization flow control—I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking (TON)*,7(6), 861-874.

Mellouk, A., Hoceini, S., & Zeadally, S. (2013). A state-dependent time evolving multi-constraint routing algorithm. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(1), 6.

Park, K. W., & Park, K. H. (2011). ACCENT: Cognitive cryptography plugged compression for SSL/TLS-based cloud computing services. *ACM Transactions on Internet Technology (TOIT)*, 11(2), 7.

Preciado, V. M., & Jadbabaie, A. (2010). Moment-based spectral analysis of large-scale networks using local structural information.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., & Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3), 630-643.

Serpanos, D. N., Mountrouidou, P., & Gamvrili, M. (2004). Evaluation of hardware and software schedulers for embedded switches. *ACM Transactions on Embedded Computing Systems (TECS)*, 3(4), 736-759.

Shue, C. A., & Kalafut, A. J. (2013). Resolvers Revealed: Characterizing DNS Resolvers and their Clients. *ACM Transactions on Internet Technology (TOIT)*, *12*(4), 14.

Smaragdakis, G., Laoutaris, N., Lekakis, V., Bestavros, A., Byers, J. W., & Roussopoulos, M. (2011). Selfish overlay network creation and maintenance. *IEEE/ACM Transactions on Networking (TON)*, *19*(6), 1624-1637.

Wu, H., Feng, Z., Guo, C., & Zhang, Y. (2010, November). ICTCP: Incast Congestion Control for TCP in data center networks. In *Proceedings of the 6th International Conference* (p. 13). ACM.

Security and privacy

Ali, M. Q., Al-Shaer, E., Khan, H., & Khayam, S. A. (2013). Automated anomaly detector adaptation using adaptive threshold tuning. *ACM Transactions on Information and System Security (TISSEC)*, 15(4), 17.

Basin, D., Jugé, V., Klaedtke, F., & Zălinescu, E. (2013). Enforceable security policies revisited. *ACM Transactions on Information and System Security (TISSEC)*, 16(1), 3.

Becchi, M., & Crowley, P. (2013). A-DFA: A Time-and Space-Efficient DFA Compression Algorithm for Fast Regular Expression Evaluation. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(1), 4.

Becker, M. Y., & Nanz, S. (2010). A logic for state-modifying authorization policies. *ACM transactions on information and system security (TISSEC)*, *13*(3), 20.

Bhargavan, K., Fournet, C., Corin, R., & Zălinescu, E. (2012). Verified cryptographic implementations for tls. *ACM Transactions on Information and System Security (TISSEC)*, *15*(1), 3.

Cabuk, S., Brodley, C. E., & Shields, C. (2009). IP covert channel detection. *ACM Transactions on Information and System Security (TISSEC)*, 12(4), 22.

Camenisch, J., & Groß, T. (2012). Efficient attributes for anonymous credentials. *ACM Transactions on Information and System Security (TISSEC)*, *15*(1), 4.

Ciriani, V., Vimercati, S. D. C. D., Foresti, S., Jajodia, S., Paraboschi, S., & Samarati, P. (2010). Combining fragmentation and encryption to protect privacy in data storage. *ACM Transactions on Information and System Security (TISSEC)*, *13*(3), 22.

Cobb, W. E., Baldwin, R. O., & Laspe, E. D. (2013). Leakage Mapping: A Systematic Methodology for Assessing the Side-Channel Information Leakage of Cryptographic Implementations. *ACM Transactions on Information and System Security (TISSEC)*, 16(1), 2.

Crampton, J., Gutin, G., & Yeo, A. (2013). On the Parameterized Complexity and Kernelization of the Workflow Satisfiability Problem. *ACM Transactions on Information and System Security (TISSEC)*, 16(1), 4.

Ferraiolo, D. F., Sandhu, R., Gavrila, S., Kuhn, D. R., & Chandramouli, R. (2001). Proposed NIST standard for role-based access control. *ACM Transactions on Information and System Security (TISSEC)*, 4(3), 224-274.

Manshaei, M., Zhu, Q., Alpcan, T., Basar, T., & Hubaux, J. P. (2011). Game theory meets network security and privacy. *ACM transaction on Computational Logic*, 5.

Mittal, P., & Borisov, N. (2012). Information leaks in structured peer-to-peer anonymous communication systems. *ACM Transactions on Information and System Security (TISSEC)*, *15*(1), 5.

Pek, G., An, L. B., & Ath, B. A. B. (2013). A Survey of Security Issues in Hardware Virtualization. *ACM Computing Surveys (CSUR)*, 45(3), 40-73.

Roemer, R., Buchanan, E., Shacham, H., & Savage, S. (2012). Return-oriented programming: Systems, languages, and applications. *ACM Transactions on Information and System Security (TISSEC)*, 15(1), 2.

Software and its engineering

Anvik, J., & Murphy, G. C. (2011). Reducing the effort of bug report triage: Recommenders for development-oriented decisions. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *20*(3), 10.

Banerjee, A., Naumann, D. A., & Rosenberg, S. (2008). Regional logic for local reasoning about global invariants. In *ECOOP 2008–Object-Oriented Programming* (pp. 387-411). Springer Berlin Heidelberg.

Barthe, G., Rezk, T., Russo, A., & Sabelfeld, A. (2010). Security of multithreaded programs by compilation. *ACM Transactions on Information and System Security (TISSEC)*, *13*(3), 21.

Chen, R., & Chen, H. (2013). Tiled-MapReduce: Efficient and Flexible MapReduce Processing on Multicore with Tiling. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(1), 3.

Dionisio, J., BURNS, W., & GILBERT, R. (2011). 3D Virtual Worlds and the Metaverse: Current Status and Future Possibilities. *Unpublished manuscript (under review, ACM Computing Surveys.)*.

Elmasry, A., & Hammad, A. (2009). Inversion-sensitive sorting algorithms in practice. *Journal of Experimental Algorithmics (JEA)*, 13, 11.

Lama, P., & Zhou, X. (2013). Autonomic Provisioning with Self-Adaptive Neural Fuzzy Control for Percentile-Based Delay Guarantee. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(2), 9.

Liu, C. L., & Layland, J. W. (1973). Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM (JACM)*, 20(1), 46-61.

Lucia, A. D., Fasano, F., Oliveto, R., & Tortora, G. (2007). Recovering traceability links in software artifact management systems using information retrieval methods. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 16(4), 13.

Mattsson, A., Fitzgerald, B., Lundell, B., & Lings, B. (2012). An Approach for Modeling Architectural Design Rules in UML and its Application to Embedded Software. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(2), 10.

Nandivada, V. K., Shirako, J., Zhao, J., & Sarkar, V. (2013). A Transformation Framework for Optimizing Task-Parallel Programs. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 35(1), 3.

Philippaerts, P., Younan, Y., Muylle, S., Piessens, F., Lachmund, S., & Walter, T. (2013). CPM:

Masking code pointers to prevent code injection attacks. *ACM Transactions on Information and System Security (TISSEC)*, 16(1), 1.

Robinson, W. N., & Ding, Y. (2010). A survey of customization support in agent-based business process simulation tools. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *20*(3), 14.

Stanier, J., & Watson, D. (2013). Intermediate representations in imperative compilers: A survey. *ACM Computing Surveys (CSUR)*, 45(3), 26.

Verdoolaege, S., Carlos Juega, J., Cohen, A., Ignacio Gómez, J., Tenllado, C., & Catthoor, F. (2013). Polyhedral parallel code generation for CUDA. *ACM Transactions on Architecture and Code Optimization (TACO)*, 9(4), 54.

Theory of computation

Austrin, P., & Håstad, J. (2013). On the usefulness of predicates. *ACM Transactions on Computation Theory (TOCT)*, 5(1), 1.

Barsky, M., Stege, U., Thomo, A., & Upton, C. (2008). A graph approach to the threshold all-against-all substring matching problem. *Journal of Experimental Algorithmics (JEA)*, 12, 1-10.

Beyersdorff, O., Datta, S., Mahajan, M., Scharfenberger-Fabian, G., Sreenivasaiah, K., Thomas, M., & Vollmer, H. (2011). Verifying proofs in constant depth. In *Mathematical Foundations of Computer Science 2011* (pp. 84-95). Springer Berlin Heidelberg.

Blume, L., Easley, D., Kleinberg, J., Kleinberg, R., & Tardos, É. (2013). Network Formation in the Presence of Contagious Risk. *ACM Transactions on Economics and Computation*, *1*(2), 6.

Bogdanov, A., Kawachi, A., & Tanaka, H. (2011). Hard functions for low-degree polynomials over prime fields. In *Mathematical Foundations of Computer Science 2011* (pp. 120-131). Springer Berlin Heidelberg.

Cederman, D., & Tsigas, P. (2009). Gpu-quicksort: A practical quicksort algorithm for graphics processors. *Journal of Experimental Algorithmics (JEA)*,14, 4.

Cygan, M., Pilipczuk, M., Pilipczuk, M., & Wojtaszczyk, J. O. (2012). On multiway cut parameterized above lower bounds. In *Parameterized and Exact Computation* (pp. 1-12). Springer Berlin Heidelberg.

Daskalakis, C. (2013). On the complexity of approximating a Nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3), 23.

Drmota, M., & Szpankowski, W. (2011, January). A master theorem for discrete divide and conquer recurrences. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 342-361). SIAM.

Ferraro-Petrillo, U., Grandoni, F., & Italiano, G. F. (2013). Data structures resilient to memory faults: An experimental study of dictionaries. *Journal of Experimental Algorithmics (JEA)*, 18(1), 1-6.

Gawrychowski, P. (2013). Optimal pattern matching in LZW compressed strings. *ACM Transactions on Algorithms (TALG)*, 9(3), 25.

Gradwohl, R., Livne, N., & Rosen, A. (2013). Sequential rationality in cryptographic protocols. *ACM Transactions on Economics and Computation*, *1*(1), 2.

Haghpanah, N., Immorlica, N., Mirrokni, V., & Munagala, K. (2013). Optimal auctions with positive network externalities. *ACM Transactions on Economics and Computation*, *1*(2), 13.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604-632.

Kobayashi, N. (2013). Model Checking Higher-Order Programs. *Journal of the ACM (JACM)*, 60(3), 20.

Ron, D., & Tsur, G. (2011). On approximating the number of relevant variables in a function. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (pp. 676-687). Springer Berlin Heidelberg.

Ševčík, J., Vafeiadis, V., Zappa Nardelli, F., Jagannathan, S., & Sewell, P. (2013). CompCertTSO: A verified compiler for relaxed-memory concurrency. *Journal of the ACM (JACM)*, 60(3), 22.

Williams, R. (2010). Alternation-trading proofs, linear programming, and lower bounds. *arXiv preprint arXiv:1001.0746*.

Appendix E: Test computer science corpus bibliography

Bao, V. N. Q., Thanh, T. T., Nguyen, T. D., & Vu, T. D. (2013). Spectrum Sharing-based Multi-hop Decode-and-Forward Relay Networks under Interference Constraints: Performance Analysis and Relay Position Optimization. *arXiv preprint arXiv:1301.0384*.

Ben-Porat, U., Bremler-Barr, A., & Levy, H. (2013). Vulnerability of network mechanisms to sophisticated DDoS attacks. *IEEE Transactions on Computers*, 62, 5.

Bjørner, D. (2006). *Software Engineering 2: Specification of systems and languages* (Vol. 2). Berlin: Springer.

Chang, C., Yang, C., Chang, Y., & Kuo, T. (2013). Booting Time Minimization for Real-Time Embedded Systems with Non-Volatile Memory. *IEEE Transactions on Computers*, 1.

Chen, Z., Sinha, A., & Schaumont, P. (2013). Using Virtual Secure Circuit to Protect Embedded Software from Side-Channel Attacks. *IEEE Transactions on Computers*, 62, 1.

Fanelli, M., Foschini, L., Corradi, A., & Boukerche, A. (2012)Self-Adaptive Context Data Management in Large-Scale Mobile Systems. *IEEE Transactions on Computers*, 1.

Han, J., Susilo, W., & Mu, Y. (2013). Identity-Based Secure Distributed Data Storage Schemes. *IEEE Transactions on Computers*, 1.

Hsiu, P., Hsieh, C., Lee, D., & Kuo, T. (2012). Multi-Layer Bus Optimization for Real-Time Embedded Systems. *IEEE Transactions on Computers*, 61, 11.

Katsikogiannis, G., Mitropoulos, S., & Douligeris, C. (2013). Policy-based QoS management for SLA-driven adaptive routing. *Communications and Networks, Journal of*, *15*(3), 301-311.

Lai, S., & Ravindran, B. (2013). Least-latency routing over time-dependent wireless sensor networks. *IEEE Transactions on Computers*, 62, 5.

Le, H., & Prasanna, V. (2013). A memory-efficient and modular approach for large-scale string pattern matching. *IEEE Transactions on Computers*, 62, 5.

Lee, J., & Lee, S. H. (2013). Low dimensional multiuser detection exploiting low user activity. *Communications and Networks, Journal of*, *15*(3), 283-291.

Lin, T. Y. (Ed.). (2006). *Foundations and novel approaches in data mining* (Vol. 9). Springer.

Luong, T., Melab, N., & Talbi, E. (2013). *GPU computing for parallel local search metaheuristic algorithms. IEEE Transactions on Computers*, 62, 1.

Papagianni, C., Leivadeas, A., Papavassiliou, S., Maglaris, V., & Monje, A. (2013). On the optimal allocation of virtual resources in cloud computing networks. *IEEE Transactions on Computers*, 62, 6.

Paulson, L. C. (2000). Foundations of Computer Science.

Peng, J., Hong, P., & Xue, K. (2013). Performance analysis of switching strategy in LTE-A heterogeneous networks. *Communications and Networks, Journal of*, *15*(3), 292-300.

Rahulamathavan, Y., Phan, R., Chambers, J., & Parish, D. (2012). Facial Expression Recognition in the Encrypted Domain based on Local Fisher Discriminant Analysis. *IEEE Transactions on Affective Computing*, 4, 1.

Sedgewick, R., & Wayne, K. (2004). Introduction to Computer Science. *Online textbook available at www. cs. princeton. edu/introcs/home*.

Villas, L., Boukerche, A., Ramos Filho, H., Oliveira, H., Araujo, R., & Loureiro, A. (2013). DRINA: a lightweight and reliable routing approach for in-network aggregation in wireless sensor networks. *IEEE Transactions on Computers*, 62, 4.

Wang, C., Wang, Q., Ren, K., & Lou, W. (2010, March). Privacy-preserving public auditing for data storage security in cloud computing. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1-9). IEEE.

Wang, S., Liu, Z., Wang, Z., Wu, G., Shen, P., He, S., & Wang, X. (2012). Analyses of a Multi-modal Spontaneous Facial Expression Database. *IEEE Transactions on Affective Computing*, 4, 1.

Yoshida, P. G. N. CONCUR 2004–Concurrency Theory.

Appendix F: Test fiction corpus bibliography

Barrie, J. M. (2008). *Peter Pan and Wendy*. 1991. *Project Gutenberg*, *28*.

Baum, L. F. (1985). *The wonderful wizard of Oz*. Castrovilli Giuseppe.

Brontë, E. (2001). *Wuthering heights*. Broadview Press.

Doyle, A. C. (2004). *The hound of the Baskervilles*. Penguin UK.

Dickens, C. (2000). *A Tale of Two Cities [1859]*. Gawthorn.

Alexandre Dumas, P. (1816). The Count of Monte Cristo.

Grahame, K. (2010). *The wind in the willows*. Oxford University Press.

Hugo, V. (1862). *Les misérables* (Vol. 5). Lassalle.

Joyce, J. (1924). *Ulysses*. Editions Artisan Devereaux.

Kafka, F. (2008). *Metamorphosis and other stories*. Penguin.

Kant, I., Guyer, P., & Wood, A. W. (Eds.). (1998). *Critique of pure reason*. Cambridge University Press.

Kipling, R. (1998). *Just so stories for little children*. Oxford University Press.

London, J. (2000). The Call of the Wild. 1903. *Online. Wiretap.[gopher://wiretap. area. com/00/Library/Classic/] Downloaded January*, *8*, 2000.

Melville, H. (1988). Moby-Dick or, The Whale. 1851. *Evanston and Chicago IL: Northwestern UP.*

Nietzsche, F. (2002). *Nietzsche: Beyond Good and Evil: Prelude to a Philosophy of the Future*. Cambridge University Press.

Plato. (2012). *The republic*. Interactive Media.

Rand, A. (2010). *Anthem*. MobileReference.

Shaw, G. B., & Loewe, F. (1975). *Pygmalion: A romance in five acts* (Vol. 2476). Penguin.

Shelley, M. W. (2008). *Frankenstein, Or, The Modern Prometheus [1818]*. Engage Books.

Stevenson, R. L. (1985). Treasure Island. 1883. *The Thistle Edition of the Works of Robert Louis Stevenson*.

Swift, J. (1844). *Gulliver's travels into several remote nations of the world*. B. Tauchnitz.

Tolstoy, L. (2010). *War and peace*. Digireads. com.

Twain, M. (2003). *Adventures of Huckleberry Finn*. Univ of California Press.

Wells, H. G. (2001). *The time machine*. Broadview Press.

Wilde, O. (1956). *The importance of being earnest: a trivial comedy for serious people*. R. Merle (Ed.). Methuen.

Wodehouse, P. G. (2010). *My Man Jeeves*. MobileReference.

References:

Adolphs, S. & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24, 425-438.

Aichah, T. (2012). *Creating a law word list and a law-specific expression list.* (Unpublished MA dissertation), Swansea University, Swansea.

Anthony, L. (2002). AntConc [Computer software]. Retrieved June 1st, 2013, from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2008). AntWordProfiler [Computer software]. Retrieved June 1st, 2013, from http://www.antlab.sci.waseda.ac.jp/software.html

Atkins, S., Clear, J., and Ostler, N. (1992) Corpus Design Criteria. *Literary Linguistic Computing*, 7, 1-16.

Bauer, L. & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6, 3, 253-279.

Biber, D. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Carroll, J.B., Davies P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. Boston, MA: Houghton Mifflin.

Chung, T. M., & Nation, P. (2003). Technical Vocabulary in Specialised Texts. *Reading in a Foreign Language*, 15(2), 103-116.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*. 11(3), 38–63.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Amsterdam, the Netherlands: John Benjamins.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*, 213-239.

Coxhead, A. (2011). The Academic Word List 10 Years On: Research and Teaching Implications. *TESOL Quarterly*, 45 (2), 355-362.

Coxhead, A., & Hirsh, D. (2007). A Pilot Science-Specific Word List. *Revue Française de linguistique appliquée*, 12(2), 65-78.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, 16(2), 35–52.

Daintith, J. & Wright, E. (2008). *Oxford dictionary of computing* (6th ed.). New York: Oxford University Press.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for

academic purposes. *English for Specific Purposes*, 28, 157-169.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL*, *6(2)*, 213–231.

Ermann, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29-62.

Faucett, L., Palmer, H., Thorndike, E. L., & West, M. (1936). *Interim report on vocabulary selection*. London: P.S. King and Son, Ltd.

Fraser, S. (2009). Technical vocabulary and collocational behaviour in a specialised corpus.In M. Edwardes (Eds.), *Language, Learning, and Context*. Paper presented at The Proceedings of the 42nd Annual Meeting of the British Association for Applied Linguistics, Newcastle University, Newcastle, 3-5 September (pp.11-48). London: Scitsiugnil Press.

Fromkin, V. (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Gilner, L., & Morales, F. (2008). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguistics*, 41–58.

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language*, 23 (1), 65-83.

Hirsch, D. (2004). *A Functional Representation of Academic Vocabulary*. (Unpublished PhD thesis), Victoria University of Wellington.

Hirsh, D. & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8, 689-696.

Hu, M. & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403-430.

Hyland, K., & Tse, P. (2007). Is There an "Academic Vocabulary?". *TESOL Quarterly*, 41(2), 235-253.

Jordan, R., R. (2002). The Growth of EAP in Britain. *Journal of English for Academic Purposes*, 1, 69-78.

Konstantakis, N. (2007). Creating a business word list for teaching English. *Estudios de lingüística inglesa aplicada*, 7, 79-102.

Konstantakis, N. (2010). *Constructing a word list for the domain of business.* (Unpublished PhD thesis), Swansea University, Swansea.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special Language: From Human Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

Laufer, B. (1990). Ease and Difficulty in Vocabulary Learning: Some teaching Implications.

Foreign Language Annals, 23, 147-155.

Laufer, B. & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22, 15–30.

Levelt, W., J., M. (1989). *Speaking: from Intention to Articulation*. Massachusetts: MIT Press.

Li, Y., & Qian, D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System*, 38, 402–411. doi:10.1016/j.system.2010.06.015.

Liu, N & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC journal*, 16(1), 33-42.

Martınez, I., Beck, S., & Panza, C. (2009). Academic vocabulary in agriculture research articles. *English for Specific Purposes*, 28, 183–198. doi:10.1016/j.esp.2009.04.003.

Martinez, R. (2013). A framework for the inclusion of multi-word units in ELT. *ELT*, 67(2), 184-198.

Martinez, R., & Murphy, V. A. (2011). Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *TESOL Quarterly*, 45(2), 24.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expression List. *Linguistics*, 33(3), 22.

Milton, J. (2006). X-Lex: The Swansea Vocabulary Levels Test. In C. Coombe, P. Davidson & D. Lloyd (Eds.). *Proceedings of the 7th and 8th Current Trends in English Language Testing (CTELT) Conference*, Vol. 4 (pp.29-39). UAE: TESOL Arabia.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multiligual Matters.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12-25.

Nation, I. S. P. (2001a). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2001b). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. M. Koski, R. D. Sell & B.Warvik (Eds.), *Language, learning and literature: Studies presented to Hakan Ringbom*. Abo Akademi University, Abo: English Department Publications 4, 167–181.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How Large a Vocabulary is Needed for Reading and Listening? *The Canadian Modern Language Review*, 63(1), 59-82.

Nation, I. S. P. (2008). *Teaching vocabulary : strategies and techniques*. Boston, MA: Heinle.

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. Retrieved June 1st, 2013, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I.S.P., & Heatley, A. (2002). Range [Computer software]. Retrieved June 1st, 2013, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*, 35–41.

Nelson, M. (2000) *A Corpus-based study of Business English and Business English Teaching Materials*. (Unpublished PhD Thesis). University of Manchester, Manchester.

O'Keefe, A., McCarthy, M. and Carter, R. (2007) *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press.

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC*, *5(2)*, 69–84.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary.* Basingstoke, England: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95 (1), 26-43.

Schmitt, N. & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, doi:10.1017/S0261444812000018.

Schonell, F.J., Middleton, I.G., & Shaw, B.A. (1956). *A Study of the Oral Vocabulary of Adults*. Brisbane: University of Queensland Press.

Sinclair, J. (2005). Corpus and text-basic principles. *Developing Linguistic Corpora: a Guide to Good Practice. Oxbow Books, Oxford*.

Verlinde, S., & Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learner's dictionary. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 594- 598). Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language.

Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28(1), 33–41. doi:10.1016/j.esp.2008.08.003.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, 27, 442-458.

Wang, K. M. T. & Nation, P. (2004). Word meaning in academic English: Homography in the academic word list. *Applied Linguistics*, 25, 291-314.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12, 309-323.

Ward, J. (2007). Collocation and technicality in EAP Engineering. *Journal of English for Academic Purposes*, 6, 18-35.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press

Zipf, G. K. (1935). *The psycho-biology of language.* New York: Houghton Mifflin.