Scientometrics
https://doi.org/10.1007/s11192-018-2778-7

An assessment into the characteristics of award winning papers at CHI

Omar Mubin[1] · Dhaval Tejlavwala[1] · Mudassar Arsalan[1] · Muneeb Ahmad[1] · Simeon Simoff[1]

Received: 2 March 2018

Abstract The overall readability of CHI publications is not known. In addition, little is understood about what lexical or demographic characteristics are unique to award winning papers at CHI and if they are significantly different from non award winning papers. We therefore carry out an exploration and assessment into the readability metrics as well as a meta analysis of 382 full papers and 54 notes from the 2014, 2015, 2016 and 2017 editions at CHI. Our results illustrate that notes did not have any significant trends whatsoever. On the other hand, award winning full papers were shown to have lower readability as compared to non award winning full papers. The type of research contribution played an important role; such that award winning full papers were significantly more likely to have a theoretical contribution as compared to non award winning full papers and full papers that presented an artifact as their contribution were more readable than other full papers. Our demographic analysis of authors indicated that the experience of authors nor their region of affiliation were not associated with the likelihood of their full paper being awarded. The experience of authors did not effect the overall readability of full papers however the region of affiliation did have a significant influence on the overall readability of full papers. In conclusion, we speculate on our obtained results through linkages with prior work in readability analysis.

Keywords CHI conference · Best papers · Readability · Special mention

Introduction

The ACM conference on Human Factors in Computing Systems (or colloquially known as the CHI conference) is one of the more renowned conferences in the area of Computer Science (Bartneck and Hu 2009). As such it is the flagship conference of the discipline of

✉ Omar Mubin
o.mubin@westernsydney.edu.au

[1] Western Sydney University, Sydney, Australia

Published online: 19 May 2018

Human Computer Interaction and is perceived to have high impact, even in comparison to some journals (Mubin et al 2017). In recent times, the typical acceptance rate of the conference (main track only) fluctuates between 20 and 25% and therefore it is an honour for researchers to have their work accepted for presentation at CHI. It is even more prestigious for authors of either full papers or notes to be nominated within the "Best of CHI" category. Specialized committees within the main Program Committee are responsible for choosing which main track papers are to be awarded at the CHI conference. The CHI conference hands out the best paper nomination (awarded to 1% of submitted papers) and an honourable or special mention nomination (awarded to 5% of submitted papers). In most instances, these specialized committees discuss nominations at the time of the main Program Committee (PC) meeting whilst deciding on the acceptance of papers. With the recent move towards virtual PC meetings, program chairs are cognizant of the added burden of selecting best papers in an already hectic schedule and have expressed interest in devising an automated or software based voting system (Young 2016). In some editions of CHI (Best of CHI 2011 Award Winners 2011), reviewers are also asked to nominate if papers that they review are deemed worthy of being awarded. However, to the best of our knowledge, the exact selection criteria employed by the specialized committees is not made public. Therefore, to an outsider the entire process may seem to encased within a black box and has hence been subject to criticism (Bartneck and Hu 2009). The main critique was that award winning papers were shown to have similar impact (or citations) as non award winning papers at CHI. Therefore in lieu of deriving certain characteristics of awarded papers in order to provide our community an opportunity to reflect on the choice of nominated best papers and to possibly reveal a bit more about the awarding criteria, we set out to determine the lexical and demographic features of CHI best papers in comparison to non awarded papers at CHI and their associated impact on readability.

It is logical to expect that the specialized committees within the PC would probably look at review ratings, topical relevance, novelty and readability of ideas and writing while adjudicating best paper and special mention awards. Whilst review ratings are not publicly available and topical relevance is subjective, the importance of readability stands out as a key aspect in academic science (Mick 2005; Zimmerman 1989). Particularly in multidisciplinary areas such as Human Computer Interaction, being able to reach out by producing readable papers to a wide audience is imperative, all of whom may have completely distinct research backgrounds. Therefore we could be led to believe that award winning papers would most likely be more readable than their compatriot non award winning papers and hence readability may well be used as a key prerequisite by the specialized committees. Readability is a commonly used metric in cognitive, literary and other sciences and is used to refer to the "ease of understanding or comprehension based on style of writing" of a piece of text (Lee and French 2011). A large variety of readability measures are available each having their own pros and cons (DuBay 2004) and different computation techniques (for example, counting number of characters or counting number of syllables in a word to determine if the word is difficult). Prior work includes a number of studies analyzing the readability of journal papers (Dolnicar and Chapple 2015), variations in readability depending on the number of authors (Hartley and Cabanac 2016), specific sections of a paper (Hartley et al 2003); such as abstracts (Kitchenham et al 2008; Guerini et al 2012), specific authors/institutions (Gazni 2011) or yearly trends (Plavén-Sigray et al 2017). Whilst a number of studies have explored the impact of award winning papers through generated citations (Coupé 2013; Wainer et al 2015; Lee et al 2003) or subsequent publications (Mittal and Gupta 2011), to the best of our knowledge there is only one recent study (Sawyer et al 2008) which attempts to compare the readability of award winning

papers against non award winning ones. In their study, Sawyer et.al compared the readability of 162 papers (81 award winning) from four top Marketing journals. Their main result was that award winning papers were easier to read. They also deduced that longer sentences and longer words had an impact on readability. Further the association between the topic of paper and the likelihood of being awarded was not clear. Lastly in their research they did not immediately establish the characteristics of award winning papers. We attempt to establish our unique contribution through a number of key advancements in comparison to the mentioned study. Firstly, papers from the CHI conference have not been exposed to a readability analysis. Secondly, our attempt is one of the first that focuses on studying the readability of the entire full text of papers from conferences from the area of Computing, given that most of the prior work has computed readability on the basis of subsections within conference papers. Thirdly, we employ a larger sample of papers and attempt to associate a number of demographic attributes (such as topical content, author affiliation) of papers with their readability measures. Therefore, we believe that we not only contribute towards Scientometrics and Infometrics literature but also allow the CHI community to reflect inwards on the overall readability of their papers and on their choice of awarded papers.

In order to assess the readability characteristics of the CHI conference we aimed to carry out a quantitative analysis from a lexical and demographic perspective. Specifically we aimed to understand the readability of CHI papers in general and if as per expectation and extending from prior work (Sawyer et al 2008), award winning papers at CHI would be easier to read than papers which were accepted but not awarded. In addition to readability and lexical metrics we also aimed to investigate the potential impact of various meta level variables (such as the number of authors, number of references, research contribution) on readability. Prior research (Sawyer et al 2008) also indicates that for specific disciplines, paper readability is contingent on the content or main methodology depicted in the article. We also wished to ascertain if certain demographic features of a paper (such as the dominating region of affiliation or the research experience of authors) would have an effect on not only the readability of a paper but also the likelihood of the paper being awarded. It would be logical to assume that experienced authors would have the most advanced literacy skills to be able to produce articles that were easier to read; this is also indicated by previous studies in the field (Sawyer et al 2008). Furthermore, specifically addressing the case of the CHI conference; the CHI conference is generally thought of a conference that is dominated by North American attendees and researchers (Bartneck and Hu 2009). We aimed to ascertain if this influence would also permeate to the likelihood of attaining award winning papers. In summary, we focused on coding the experience of authors, their region of affiliation and the research contribution of papers in addition to the basic meta level categories. We conclude this section by summarizing a succinct list of research questions.

- *RQ1* Are award winning papers at CHI easier to read than their non award winning compatriots?
- *RQ2* What impact does the type of research described in CHI papers have on both the likelihood of them being awarded and on their readability?
- *RQ3* What impact do the demographic attributes of the authors of a CHI paper (dominant regional affiliation, research experience) have on both the likelihood of the papers being awarded and on their readability?

- *RQ4* What impact do the meta attributes of a CHI paper (number of references, authors, tables and images) have on both the likelihood of the papers being awarded and on their readability?

Method

Our methodology was based on textual analysis of a corpus of CHI papers using the Python programming language (see "Appendix" for entire Python code). Thereafter using statistical significance testing in SPSS we aimed to determine whether particular readability parameters or other meta level attributes were different across both award winning and non award winning papers. We selected CHI main track (Full and Notes) papers from the years 2014, 2015, 2016 and 2017 as the primary component of our sample. We believed that focusing on the CHI proceedings from the most recent years would allow us to delve into the latest trends and characteristics of CHI best papers.

Data collection and processing

Via the ACM digital library, we first collated PDF's of the entire CHI conference proceedings from 2014 to 2017 and noted which papers were granted either the best paper or special mention award. As a second step we used the PyPDF2 library (pypdf2 2016) to convert the PDF's in our sample to plain text. Any text before the abstract and after and inclusive of the reference section was also discarded. That is, text extracted was abstract onwards up until the start of the bibliography section. At this juncture we noticed that not all PDF's were accurately converted to plain text; in fact only those papers which were written and edited in Microsoft Word were being converted to readable text. Papers written in LaTeX or other document editing tools were not being converted properly by the Python library and resulted in junk output. Therefore, our final sample of cleanly converted papers amounted to 975 across the four years of the CHI conference (145 Notes and 830 full papers), which equated to approximately 46% of the entire main track proceedings (a total of 2116 papers were accepted across the four years). At this point, we also merged the two types of award winning papers (best papers and special mention) into one category in order to simplify the ensuing data analysis. This gave us a total of 218 award winning papers across the sample of 975 papers (191 full papers and 27 notes). In order to have a comparable proportion of both award winning and non award winning we randomly selected an equal amount of non award winning papers against the retrieved award winning papers; separately for both full papers and notes. This step gave us a final sample of 436 papers comprising of 382 full papers from the years 2014–2017 (including 191 award winning papers) and 54 short papers (or Notes) from the years 2014–2016 (which included 27 award winning papers). Unfortunately, our Python script was unable to extract the text for not even a single award winning note from CHI 2017.

The entire coverage of award winning papers (218 papers) in our sample from the editions CHI 2014 to 2017 amounted to roughly 50% of the total set of award winning papers (434 papers)—(see Table 1), providing us with sufficient confidence to pursue significance testing on this data. A total of 539 non award winning papers from our sample were hence ignored from hereon in, in lieu of achieving an equal proportion of both award winning and non award winning papers.

Table 1 Our coverage of award and non award winning papers at CHI 2014–CHI 2017

CHI meta data					Our sample size	
Year	Total papers accepted	Non awarded papers	Best papers	Honourable mention	Non awarded papers	Awarded papers
2014	465	360	21	84	72	79
2015	486	367	21	98	60	63
2016	565	476	20	69	68	58
2017	600	479	24	97	18	18

Measurements

Primarily three levels of data were extracted from each paper using either Python libraries or manual extraction techniques. Firstly, the meta level information associated with each paper was recorded, which included: the number of authors, the number of images, the number of tables and the number of references. In addition for every paper, we also recorded the popular region of affiliation (or the most dominant) among all authors, which was coded as: South Africa, East Asia, North America, Northern Europe, Oceania, South-East Asia, Southern Europe, Western Europe, Middle East and South Asia. Any regions outside of the list mentioned prior were not seen in our sample. Furthermore, for every paper in our sample, we noted the total experience of the authors in terms of absolute number of publications as well as the duration for which the authors have been research active; up until February 2018. This was noted from the authors tab in ACM digital library for every paper. The fact that we were unable to determine both measures at time of when the paper was published should be considered as a limitation of our analysis. Lastly, using the taxonomy presented in (Wobbrock and Kientz 2016), we also coded the research contribution of every paper into 7 possible codes. As such the taxonomy presented is ideal for our analysis because it is specific to the field of Human Computer Interaction. Papers could either have an empirical, artifact, methodological, theoretical, dataset based, survey or opinionated contribution. Two authors shared the task of coding the research contribution of 30 randomly selected papers from the sample of 436. Moderate agreement ($\kappa < 0.6$) was found and ambiguities were resolved after consultation. The first author then finished the coding of the entire sample.

Secondly, using the nltk package (Bird et al 2009) a number of lexical features were extracted, namely: word count, sentence count, average words per sentence and the number of unique words (or also known as tokens). Thirdly, using the textstat Python library (Plavén-Sigray et al 2017), a suite of popular, reliable and widely utilized readability indexes were computed. These included: Flesch-Reading Ease, Flesch–Kincaid Grade, Gunning Fog index, SMOG index, automated readability index, Coleman Liau index, Linsear write formula, Date Chal Readability Score and the number of difficult words in the paper (which are words other than the 3000 most common words in the English language). At this stage we do not define each readability index as a number of prior research articles have already discussed the pros and cons of each metric and their different computation techniques (Lee and French 2011). Extending from (Sawyer et al 2008) we also computed an *average readability grade* which was the average of the following four metrics: Flesch–Kincaid Grade, Gunning Fog index, SMOG index and automated

readability index. It is worth mentioning that other than the Flesch-Reading Ease, for all other indexes the lower the numeric measurement the more readable the document is. The Flesch-Reading Ease index is measured on a scale of 0–100 (30 or below is difficult and 70 or above is easy to read), whereas the other indices indicate the grade level of the document on the US education system (for example a readability index of 13 would correspond to the US college level or the fact that the document requires college level education).

Results

We would first like to report on some descriptive results across the entire spectrum of our data ranging from CHI2014 to 2017. We separately analyzed full papers and notes as we believed that due to large variations in page length some of the lexical parameters would be significantly different. This was achieved by simply splitting our data set in SPSS and identical tests were run for both samples. We present a table that summarizes the average readability grade (computed as the average of the 4 readability metrics as described earlier) across our entire data set (see Table 2). The associated Cronbach Alpha reliability values for the average readability grade for full papers and notes was acceptable; 0.86 and 0.91 respectively. The average readability grade for full papers and CHI notes was quite similar; 12.92 and 12.87 respectively. We also present a table with the most and least readable papers from within our sample of 346 CHI full papers and 54 CHI Notes as measured on the parameter average readability grade (see Table 3). The average word count for Notes was 3596 and for full papers was 9032. The average Flesch Reading Ease value for Notes was 55.9 whereas for Full papers it was 53.3. On average, the number of authors for full papers was 4.23 and for Notes it was 3.83. The type of research contribution of Full Papers also had an effect on the overall average readability grade—($F(5, 376) = 8.28$, $p < 0.001$). Bonferroni post hoc tests revealed that full papers that had an artifact based contribution were significantly easier to read (as per the average readability grade) than empirical or theoretical papers ($p < 0.001$). There were certain regional differences in readability, such that full papers predominantly written by Western European authors were more difficult to read as compared to papers predominantly written by East Asian authors ($F(7, 374) = 3.84$, $p < 0.001$); Bonferroni post hoc test $p = 0.002$; again determined on the basis of the average readability grade. North American affiliations emerged as the dominant affiliation of CHI Full papers accounting for 56% of award winning papers and 52% of non award winning papers. Western Europe was the second most dominant region accounting for almost 28% of the entire sample of 400 papers.

Table 2 Average readability grade across entire data set

Award category	Paper type	N	Average readability and SD in brackets
Non award winner	Full	191	12.90 (1.08)
	Note	27	13.07 (1.11)
Special mention	Full	155	12.69 (1.18)
	Note	22	12.92 (1.18)
Best paper	Full	36	13.30 (1.89)
	Note	5	12.64 (0.84)

Table 3 Most readable and least readable CHI papers

Paper type	Paper citation	Award winning	Average readability grade
Most readable			
Full paper	Luo, Yuexing, and Daniel Vogel. "Crossing-based selection with direct touch input." CHI 2014 (Luo and Vogel 2014)	Yes	10.40
Note	Felt, Adrienne Porter, Robert W. Reeder, Hazim Almuhimedi, and Sunny Consolvo. "Experimenting at scale with google chrome's SSL warning." CHI 2014 (Felt et al 2014)	Yes	10.94
Least readable			
Full paper	Zhang, Yunfeng, and Anthony J. Hornof. "Understanding multitasking through parallelized strategy exploration and individualized cognitive modeling." CHI 2014 (Zhang and Hornof 2014)	Yes	18.57
Note	Katan, Simon, Mick Grierson, and Rebecca Fiebrink. "Using interactive machine learning to support interface development through workshops with disabled people." CHI 2015 (Katan et al 2015)	No	15.82

In order to evaluate our main research question of testing the differences across award winning and non award winning papers at CHI, we executed two separate ANOVA's (one each for Notes and Full papers) with the category of the paper (award winning or not) as our main independent variable and our three categories of measurements (meta level information, lexical variables, readability indices) as our dependent variables. We could not find any significant results for Notes for each and every test that we ran. Linsear Write Formula was the closest to approaching significance $F(1, 51) = 3.51$, $p = 0.07$ (average for award winning papers = 11.01 and average for non award winning papers = 12.93). Due to the absence of any significant trends we do not report the specific test results on Notes. From here on in any results that are reported were generated on our sample of 382 CHI full papers.

Our second ANOVA on our sample of 382 full papers revealed 4 results that are worth mentioning—see Table 4. The number of unique words (or tokens) were significantly higher in award winning full papers as compared to non award winning full papers; $F(1, 344) = 6.99$, $p = 0.01$. The number of difficult words was also significantly higher in award winning full papers as compared to non award winning full papers; $F(1, 344) = 6.26$, $p = 0.01$. The Gunning Fog index also showcased a nearly significant trend—non award winning full papers were deemed to be more readable; $F(1, 344) = 3.27$, $p = 0.06$. Lastly, the Date Chal Readability score was also significant in favour of non award winning full papers; $F(1, 344) = 3.99$, $p = 0.05$. The effect sizes for all three significant variables was η_p^2=0.02. Subsequent association testing was done among the categorical variables. Results reported pertain only to full papers as there were no significant results for short papers (or Notes). Theoretical research contribution of full papers was shown to have a significant association with the likelihood of a paper being awarded ($\chi^2(N = 382, 5) = 10.9$, $p = 0.05$, residual = 2.4). 67% of theoretical papers (30 from a total of 45) in our sample were awarded with either a best paper or special nomination award. Full papers that had an empirical or artifact based research contribution

Table 4 Mean and standard deviations (in brackets) for all measurements across both award winning and non award winning full papers

Measurement	Non award winning papers	Award winning papers
Num of authors	4.19 (2.21)	4.26 (2.02)
Num of images	5.30 (4.24)	5.48 (4.68)
Num of tables	1.20 (1.74)	1.16 (1.80)
Num of references	38.21 (13.11)	39.62 (13.51)
Word count	8921 (1141)	9141 (1302)
Sentence count	503 (90)	507 (104)
Avg words per sentence	18.21 (2.27)	18.57 (2.41)
Num of tokens*	2297 (434)	2417 (456)
Flesch-Reading Ease	53.06 (6.69)	53.60 (7.12)
Flesch–Kincaid Grade	10.55 (1.27)	10.53 (1.37)
Gunning Fog	17.64 (1.35)	17.90 (1.43)
SMOG	10.17 (0.85)	10.07 (0.80)
Automated readability index	13.21 (1.37)	13.25 (2.07)
Coleman Liau	13.64 (1.20)	13.57 (2.22)
Linsear write formula	11.66 (3.93)	11.80 (3.88)
Date chal readability score*	7.77 (0.38)	7.85 (0.39)
Num of difficult words*	1770 (301)	1851 (328)
Average readability grade	12.90 (1.07)	12.94 (1.22)
Author experience (total years)	39.11 (24.3)	41.37 (23.21)
Author experience (total publications)	146.50 (113.4)	164.74 (113.3)

*Significant variable at the $p < 0.05$ level

were roughly equal in number (170 against 142 respectively), while there was not a single paper that we coded as having an opinion based contribution. There were 17 papers that we deemed to have a methodological contribution. Furthermore, the dominant region of affiliation for full papers did not have a significant association with the paper being ultimately awarded ($\chi^2(N = 346, 8) = 7.42, p = 0.49$). Lastly, an ANOVA test revealed that neither did the overall experience of authors of CHI full papers in our sample in terms of their total publications or active time period as a researcher had a significant effect on the likelihood of the paper being awarded an award.

We also computed bivariate correlations among some of the scale measurements from our sample of 382 CHI full papers. As expected the readability measures were highly correlated with one another. Average readability grade was not significantly correlated with experience of authors as exemplified by total number of years or total number of publications. The number of images and tables had a significant negative correlation with the average readability grade ($r = -0.31$ and -0.27). The number of authors did not have a significant correlation with any of the measurements except difficult words and date chal readability score; ($r = 0.11$, $p = 0.04$) and ($r = 0.13$, $p = 0.01$) respectively. Number of references had significant correlations with a number of lexical features all on a $p < 0.001$ level; including difficult words—$r = 0.58$, word count—$r = 0.56$, sentence count—$r = 0.78$, tokens—$r = 0.52$ and words per sentence—$r = -0.50$.

Discussion

Our readability analysis of CHI main track papers in general and a comparison between award winning and non award winning papers has revealed some interesting insights. Firstly, the overall readability of CHI papers was found to be much easier in comparison to journal papers. On average the readability of CHI papers was found to be less than 13, which equates to college level education. In comparison, most other studies which have investigated the readability of journal articles have reported a higher (hence more difficult to read) readability grade, such as 15.5 (Lee and French 2011), 16.52 (Lei and Yan 2016) or 16.2 (Sawyer et al 2008). Similarly, the Flesch Reading Ease (FRE) value for CHI publications in our sample was significantly higher (hence easier to read) as compared to the FRE value for journal articles; for example in finance journals (Lee and French 2011), where a mean FRE score of 30.3 was reported.

Secondly, our readability analysis on award winning CHI papers showed that whilst the two types of Notes did not have significant readability differences, award winning full papers were emerging difficult to read as compared to non award winning full papers (*RQ1*). This transpired through not only readability indices (date chal readability score) but also via the richness of vocabulary (significantly more unique and difficult words). As a supplementary check, we also conducted an ANOVA by unwrapping the two types of awards (best paper and special mention). Yet again, we attained significant results of either of the two awarded papers over non awarded papers via lower readability and rich/dense vocabulary. This result is in contrast to the analysis reported in (Sawyer et al 2008), where award winning papers were deemed to be more readable. This trend is difficult to explain and is maybe a byproduct of a perception that difficult papers are considered to be more prestigious (Armstrong 1989) or having more impact (Lei 2016) and hence evaluated as such by the Program Committee. In our analysis, we have attempted to utilize a wide range of readability measures as it is a highly subjective construct. Although a number of indices are available, prior research (Okulicz-Kozaryn 2013) has also explored the impact of adverbs and adjectives on the readability of publications. Follow up research (Lei 2016) has criticized the use of adjectives or adverbs alone as a determinant of readability. There could be further interesting vocabulary or grammar constructs worth highlighting that influence readability of texts, such as punctuation (DuBay 2004). As a quick experiment we ran some tests on our sample of award winning and non award winning CHI full papers from 2014 to 2017 (382 papers) by extracting different kinds of punctuation. A key significant result was that the number of commas per sentence was significantly more in award winning papers as compared to simply accepted papers ($F(1, 380) = 5.01$, $p = 0.03$). In summary, whilst we have found certain unexpected trends in terms of the readability of award winning and non award winning papers, the interpretation of readability by the Program Committee and to what extent it is considered before awarding a paper is not immediately transparent.

Our results also showed that whilst full papers which had a theoretical contribution were more likely to be awarded, full papers that disseminated an artifact as their main contribution were easier to read (*RQ2*). Prior research illustrates how vocabulary, grammar constructs and consequently readability varies across disciplines (Okulicz-Kozaryn 2013). This is critical for a multi-disciplinary research area such as Human Computer Interaction. Furthermore, the systems research versus empirical research debate is a longstanding one in Human Computer Interaction (Landay 2009; Olsen Jr 2007) and roughly 82% of the full papers in our sample had an either empirical or systems/artifact based contribution. The

fact that theoretical papers were being awarded may indicate that the community is perhaps opening up to acknowledging and appreciating papers that have a theoretical impact due to their rarity and uniqueness. The fact that artifact papers were easier to read in comparison to both theoretical and empirical papers could be explained by the defined structure of such papers (to the extent that the recipe of what makes a good and typical CHI paper has been formulated (Nacke 2017) and the fact that every conference or venue is known to develop its own writing style (Alluqmani and Shamir 2018). Whereas HCI has a number of research methodologies and paradigms that would be translated into the theoretical and empirical type of papers, most artifact based papers followed a conventional design, implementation and evaluation research cycle. This consistency in format, structure and expectation in content could possibly improve the readability of such papers. Our earlier described result of Western European authors writing papers that were more difficult to read as compared to East Asian authors does indicate that readability is governed by the writing styles of authors which is likely to be associated with their region of affiliation (*RQ3*). However while the regional influence extended to overall dominance of the CHI conference (North American was the stand out performer), this was not the case for the likelihood of papers being awarded (*RQ3*).

Similar to (Sawyer et al 2008), the number of authors were not related with the overall readability of full papers (*RQ4*). We did observe that having more tables and images led to an overall improvement in readability (RQ4). Furthermore, unlike indications from prior work the experience of authors (total years or total publications) did not significantly correlate with overall readability (represented by the average readability grade) (*RQ3*). The number of references in full papers were positively correlated with vocabulary richness (*RQ4*). Prior research (Costas et al 2012) indicates that top performing or young researchers have a tendency to include more references as a means to show evidence or proof of their claims. Prior research also suggests how dense writing is commonly associated with a higher chance of acceptance (Lee and French 2011). This may directly correlate with the writing and referencing patterns of young researchers who are eager for important paper acceptances (for the purposes of confirming tenure or funding success).

Conclusion and future work

Significance testing on our sample of CHI papers has shown that awarded papers at the conference were observed to have lower readability. We discuss how this may be a consequence of the "bafflegab" theory (Armstrong 1989). However in general our results also show that the readability of CHI papers is much easier than journal publications. Our results also show that readability is influenced by topical content and author background but not by the research experience of authors. We also explored how certain thematic areas in papers were rare, unique and less conventional which although would have decreased their overall readability but perhaps made them more appealing to the program committee to award them. Our overall analysis was compounded and limited by the sample size, particularly in the case of CHI notes (short papers). In the future we will explore additional techniques and libraries to accurately extract raw text from PDF papers so that we can compute the readability and lexical attributes of a larger sample. At this juncture, we must reflect on our chosen methodology and data extraction technique. Papers written in LaTeX

or typesetting tools other than Microsoft Word were not extracted by our Python scripts and hence excluded from our final sample. This is a possible confound that should be addressed in future work. Prior research (Knauff and Nejasmic 2014) points out how LaTeX is a popular tool to depict mathematical or algorithmic content. Hence, one could assume that such papers may be even more difficult to read due to their technical nature and hence the overall readability of CHI may in reality be lower than what we found. The penetration of LaTeX in Computer Science is about 50% but in the area of Psychology it is next to nothing (Brischoux and Legagneux 2009). CHI is a highly multidisciplinary conference but primarily features papers from Computer Scientists and Psychology researchers as the two main attributing fields to the area. Although our data extraction rate in the first instance was 46% we may have missed out on a fair number of papers written by experts from Computer Science.

In addition, we believe that readability would also be expected to play a significant role when it comes to the acceptance of papers at CHI, however we do not have access to rejected papers from CHI at this moment to carry out this check. It would also be worthwhile to compare the readability of CHI papers against other top ranked HCI conferences or even conferences from domain areas other than Computer Science. We also believe our proposed methodology can be extended to other academic outlets such as grant applications. Lastly, we would also like to acknowledge that our analysis is constrained due to the shallow nature of readability metrics. Understanding the readability of academic articles must go beyond lexical computations and future work should utilize other measures of readability such as subjective feedback from readers and reviewers. In conclusion, it is widely acknowledged that readability is a key aspect within the communication of science and editing software would do well to provide a real time readability assessment and improvement suggestions to authors. The PC of the CHI conference should provide additional information on how papers are awarded (ideally selection criteria) so that authors and researchers can tailor their writing styles. In conclusion, the CHI conference should not implicitly encourage publications that are difficult to comprehend, particularly since the discipline of Human Computer Interaction that we belong to is highly multidisciplinary.

Appendix: Python code

Appendix A Python Code

```
# Load important libraries
import sys
import nltk
import glob
import os
import scipy
import collections
import PyPDF2 as pyPdf
import numpy as np
import sys
import warnings
import curses
import re
from os import path
from sklearn.feature_extraction.text import
    CountVectorizer
from PIL import Image
from curses.ascii import isdigit
from nltk.corpus import cmudict
from textstat.textstat import textstat

word_tokenizer = nltk.tokenize.RegexpTokenizer(r'\w+')
sentence_tokenizer = nltk.data.load('tokenizers/punkt/
    english.pickle')
data_folder = r"data/CHI2017/CHI2017_2"
files = sorted(glob.glob(os.path.join(data_folder, "*.
    pdf")))
string_for_excel = ""
counter = 1
for fn in files:
    with open(fn,'rb') as f:

        # 1. File Name

paper_number = fn[23:-4]
string_for_excel += str(paper_number) + ","

# 2. Year
published_year = "2017"
string_for_excel += published_year + ","

# Reading PDF
read_pdf = pyPdf.PdfFileReader(f)

# 3. Count number of pages
try:
    total_pages = read_pdf.getNumPages()
except:
    total_pages = 0
string_for_excel += str(total_pages) + ","

# Creating raw text
text = ""
for i in range(0,total_pages):
    text += read_pdf.getPage(i).extractText()
text = text.replace('\n', ' ')

# Generate pure text by removing title, author's
    name, institute's name and references
```

```
try:
    pure_text = text.split("ABSTRACT")
    pure_text = pure_text[1].split("REFERENCE")
    pure_text = pure_text[0]
except:
    pure_text = text

# 4. Count images
image_count = 0
image_pattern = r"s?[Figure^Fig]+[ ^  ]+([0-9])
    +\.\ "
image_count = len(re.findall(image_pattern,
    pure_text))

string_for_excel += str(image_count) + ","

# 5. Count Words in text
try:
    word_count = len(pure_text.split())
except:
    word_count = 0
string_for_excel += str(word_count) + ","

# 6. Total number of sentences
try:
    sentences = sentence_tokenizer.tokenize(
        pure_text)
    sentence_count = len(sentences)
except:
    sentence_count = 0
string_for_excel += str(sentence_count) + ","

# 7. Average words per sentence
try:
    words_per_sentence = np.array([len(
        word_tokenizer.tokenize(s)) for s in
        sentences])
    avarage_words_per_sentence = scipy.mean(
        words_per_sentence)
except:
    avarage_words_per_sentence = 0
string_for_excel += str(
    avarage_words_per_sentence) + ","

# 8. Number of letters
letters_count = 0
try:
    for char in pure_text:
        if char != " ":
            letters_count = letters_count+ 1
        else:
            continue
except:
```

```
        letters_count = 0
    string_for_excel += str(letters_count) + ","

    # 9. Unique words
    try:
        tokens = nltk.word_tokenize(pure_text.lower
            ())
        unique_words = len(set(tokens))
    except:
        tokens = 0
    string_for_excel += str(unique_words) + ","

    # 10. Average letters per word
    try:
        letters_per_word = letters_count/word_count
    except:
        letters_per_word = 0
    string_for_excel += str(letters_per_word) + ","

    # 11. Total number of syllables
    try:
        syllables_count = textstat.syllable_count(
            pure_text)
    except:
        syllables_count = 0
    string_for_excel += str(syllables_count) + ","

    # 12. Syllables per sentence
    try:
        syllables_per_sentence = syllables_count/
            sentence_count
    except:
        syllables_per_sentence = 0
    string_for_excel += str(syllables_per_sentence)
        + ","

    # 13. Questions count
    try:
        questions_count = tokens.count('?')
    except:
        questions_count = 0
    string_for_excel += str(questions_count) + ","

    # 14. Commas per sentence
    try:
        commas_per_sentence = tokens.count(',') /
            float(sentence_count)
    except:
        commas_per_sentence = 0
    string_for_excel += str(commas_per_sentence) +
        ","

    # 15. Semicolon per sentence
```

```
try:
    semicolon_per_sentence = tokens.count(';') /
        float(sentence_count)
except:
    semicolon_per_sentence = 0
string_for_excel += str(semicolon_per_sentence)
    + ","

# 16. Colon per sentence
try:
    colon_per_sentence = tokens.count(':') /
        float(sentence_count)
except:
    colon_per_sentence = 0
string_for_excel += str(colon_per_sentence) +
    ","

# First, second, third person references
first_person_counter = 0
second_person_counter = 0
third_person_counter = 0
try:
    for word in pure_text.split():
        if(word.lower() == "i" or word.lower()
            == "we"):
            first_person_counter =
                first_person_counter + 1
        elif (word.lower() == "you"):
            second_person_counter =
                second_person_counter + 1
        elif (word.lower() == "he" or word.lower
            () == "she" or word.lower() == "it"
            or word.lower() == "they"):
            third_person_counter =
                third_person_counter + 1
except:
    first_person_counter = 0
    second_person_counter = 0
    third_person_counter = 0

# 17. First person reference
try:
    first_person_reference =
        first_person_counter * 100 /
        sentence_count
except:
    first_person_reference = 0
string_for_excel += str(first_person_reference)
    + ","

# 18. Second person reference
try:
    second_person_reference =
        second_person_counter * 100 /
```

```
        sentence_count
except:
    second_person_reference = 0
string_for_excel += str(second_person_reference)
    + ","

# 19. Third person reference
try:
    third_person_reference =
        third_person_counter * 100 /
        sentence_count
except:
    third_person_reference = 0
string_for_excel += str(third_person_reference)
   + ","

# 20. No Person reference
try:
    no_person_reference = 100- (
        first_person_reference +
        second_person_reference +
        third_person_reference)
except:
    no_person_reference = 0
string_for_excel += str(no_person_reference) +
    ","

# 21. Abstract size
try:
    abstract = text.split("ABSTRACT")
    abstract = abstract[1].split("AUTHOR
        KEYWORDS")
    abstract_size = len(abstract[0].split())
except:
    abstract_size = 0
string_for_excel += str(abstract_size) + ","

# 22. Reference size
try:
    reference = text.split("REFERENCES")
    reference_size = len(reference[1].split())
except:
    reference_size = 0
string_for_excel += str(reference_size) + ","


# 23. Flesch reading ease
try:
    flesch_reading_ease = textstat.
        flesch_reading_ease(pure_text)
except:
    flesch_reading_ease = 0
string_for_excel += str(flesch_reading_ease) +
    ","
```

```
# 24. Flesch kincaid grade
try:
    flesch_kincaid_grade = textstat.
        flesch_kincaid_grade(pure_text)
except:
    flesch_kincaid_grade = 0
string_for_excel += str(flesch_kincaid_grade) +
    ","

# 25. Gunning fog
try:
    gunning_fog = textstat.gunning_fog(pure_text
        )
except:
    gunning_fog = 0
string_for_excel += str(gunning_fog) + ","

# 26. SMOG index
try:
    smog_index = textstat.smog_index(pure_text)
except:
    smog_index = 0
string_for_excel += str(smog_index) + ","

# 27. Automated readability index
try:
    automated_readability_index = textstat.
        automated_readability_index(pure_text)
except:
    automated_readability_index = 0
string_for_excel += str(
   automated_readability_index) + ","

# 28. Coleman liau index
try:
    coleman_liau_index= textstat.
        coleman_liau_index(pure_text)
except:
    coleman_liau_index = 0
string_for_excel += str(coleman_liau_index) +
    ","

# 29. Linsear write formula
try:
    linsear_write_formula = textstat.
        linsear_write_formula(pure_text)
except:
    linsear_write_formula = 0
string_for_excel += str(linsear_write_formula) +
     ","

# 30. Dale-Chall_readability_score
try:
    dale_chall_readability_score = textstat.
```

```
        dale_chall_readability_score(pure_text)
except:
    dale_chall_readability_score = 0
string_for_excel += str(
   dale_chall_readability_score) + ","

# 31. Text standard
try:
    text_standard = textstat.text_standard(
        pure_text)
except:
    text_standard = 0
string_for_excel += str(text_standard) + ","

# 32. Difficult words
try:
    difficult_words = textstat.difficult_words(
        pure_text)
except:
    difficult_words = 0
string_for_excel += str(difficult_words) + ","

# 33. Reference count
try:
    #reference = text.split("REFERENCES")
    reference_count = (len(reference[1].split
        ("[")) - 1)
except:
    reference_count = 0
string_for_excel += str(reference_count) + ","

# 34. Table count
try:
    table_count = 0
    table_pattern = r"s?Table+[ ^  ]+([0-9])+\.\
        "
    table_count = sum(1 for x in re.finditer(
        table_pattern, text))
except:
    table_count = 0
string_for_excel += str(table_count) + ","

# 35. Author count
try:
    author_count = 0
except:
    author_count = 0
string_for_excel += str(author_count) + ","

# 36. Target values
try:
    label = 2
except:
    label = 0
```

```
            string_for_excel += str(label) + ""

            string_for_excel += "\n"

            print("Counter :",counter, "File:",f)
            counter += 1

print("---")
# DATASET FEATURES (PREDICTORS)
text = "paper_number," # 1 DONE
text += "published_year," # 2 DONE
text += "page_count," # 3 DONE
text += "image_count," # 4 DONE
text += "word_count," # 5 DONE
text += "sentence_count," # 6 DONE
text += "words_per_sentence," # 7 DONE
text += "letters_count," # 8 DONE
text += "tokens," # 9 DONE
text += "letters_per_word," # 10 DONE
text += "syllables_count," # 11 DONE
text += "syllables_per_sentence," # 12 DONE
text += "questions_count," # 13 DONE
text += "commas_per_sentence," # 14 DONE
text += "semicolon_per_sentence," # 15 DONE
text += "colon_per_sentence," # 16 DONE
text += "first_person_reference," # 17 DONE
text += "second_person_reference," # 18 DONE
text += "third_person_reference," # 19 DONE
text += "no_person_reference," # 20 DONE
text += "abstract_size," # 21 DONE
text += "reference_size," # 22 DONE
text += "flesch_reading_ease," # 23 DONE
text += "flesch_kincaid_grade," # 24 DONE
text += "gunning_fog," # 25 DONE
text += "smog_index," # 26 DONE
text += "automated_readability_index," # 27 DONE
text += "coleman_liau_index," # 28 DONE
text += "linsear_write_formula," # 29 DONE
text += "dale_chall_readability_score," # 30 DONE
text += "text_standard," # 31 DONE
text += "difficult_words," # 32 DONE

text += "reference_count," # 33 NOT DONE
text += "table_count," # 34 DONE
text += "author_count," # 35 NOT DONE

text += "label" # DONE > Target values

text += "\n"
print(len(text.split(',')))
36
print("Creating CSV file")
# Open File
output_file = open("test_2017_2.csv",'w')
# Write data to file
newString = ""
```

```
newString = text + string_for_excel
#print(newString)
for r in newString:
    output_file.write(r)
output_file.close()
print("File saved!")
```

References

Alluqmani, A., & Shamir, L. (2018). Writing styles in different scientific disciplines: A data science approach. *Scientometrics*, *115*(2), 1–15.

Armstrong, J. S. (1989). Readability and prestige in scientific journals. *Journal of Information Science*, *15*, 123–124. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=668145.

Bartneck, C., & Hu, J. (2009) Scientometric analysis of the chi proceedings. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 699–708). ACM.

Best of CHI 2011 Award Winners. (2011). Best of CHI 2011 award winners. http://www.chi2011.org/program/awards.html. Accessed 15 Feb 2018.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media Inc.

Brischoux, F., & Legagneux, P. (2009). Don't format manuscripts. *The Scientist*, *23*(7), 24.

Costas, R., Leeuwen, T. N., & Bordons, M. (2012). Referencing patterns of individual researchers: Do top scientists rely on more extensive information sources? *Journal of the Association for Information Science and Technology*, *63*(12), 2433–2450.

Coupé, T. (2013). Peer review versus citations—An analysis of best paper prizes. *Research Policy*, *42*(1), 295–301.

Dolnicar, S., & Chapple, A. (2015). The readability of articles in tourism journals. *Annals of Tourism Research*, *52*, 161–166.

DuBay, W. H. (2004). The principles of readability. Online Submission. https://files.eric.ed.gov/fulltext/ED490073.pdf.

Felt, A. P., Reeder, R. W., Almuhimedi, H., & Consolvo, S. (2014). Experimenting at scale with google chrome's SSL warning. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems* (pp. 2667–2670). ACM.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science*, *37*(3), 273–281.

Guerini, M., Pepe, A., & Lepri, B. (2012). Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 475–478).

Hartley, J., & Cabanac, G. (2016). Are two authors better than one? Can writing in pairs affect the readability of academic blogs? *Scientometrics*, *109*(3), 2119–2122.

Hartley, J., Pennebaker, J., & Fox, C. (2003). Abstracts, introductions and discussions: How far do they differ in style? *Scientometrics*, *57*(3), 389–398.

Katan, S., Grierson, M., & Fiebrink, R. (2015). Using interactive machine learning to support interface development through workshops with disabled people. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 251–254). ACM.

Kitchenham, B. A., Brereton, O. P., Owen, S., Butcher, J., & Jefferies, C. (2008). Length and readability of structured software engineering abstracts. *IET software*, *2*(1), 37–45.

Knauff, M., & Nejasmic, J. (2014). An efficiency comparison of document preparation systems used in academic research and development. *PloS One*, *9*(12), e115,069.

Landay, J. (2009). I give up on CHI/UIST. Blog entry. Retrieved November 10, 2009.

Lee, J., Vicente, K., Cassano, A., & Shearer, A. (2003). Can scientific impact be judged prospectively? A bibliometric test of simonton" s model of creative productivity. *Scientometrics*, *56*(2), 223–232.

Lee, S., & French, N. (2011). The readability of academic papers in the journal of property investment & finance. *Journal of Property Investment & Finance*, *29*(6), 693–704.

Lei, L. (2016). When science meets cluttered writing: adjectives and adverbs in academia revisited. *Scientometrics*, *107*(3), 1361–1372.

Lei, L., & Yan, S. (2016). Readability and citations in information science: Evidence from abstracts and articles of four journals (2003–2012). *Scientometrics*, *108*(3), 1155–1169.
Luo, Y., & Vogel, D. (2014). Crossing-based selection with direct touch input. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2627–2636). ACM.
Mick, D. (2005). Inklings: From mind to page in consumer research. In *Association of consumer research newsletter* (pp. 1–3).
Mittal, H., & Gupta, P. (2011). Fate of award winning papers at annual conference of Indian academy of pediatrics: A 13 years experience. *Indian Pediatrics*, *48*, 818–819. https://www.indianpediatrics.net/oct2011/oct-818-819.htm.
Mubin, O., Al Mahmud, A., & Ahmad, M. (2017). HCI down under: Reflecting on a decade of the OzCHI conference. *Scientometrics*, *112*(1), 367–382.
Nacke, L. E. (2017). How to write and review chi papers. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (pp. 1228–1231). ACM.
Okulicz-Kozaryn, A. (2013). Cluttered writing: Adjectives and adverbs in academia. *Scientometrics*, *96*(3), 679–681.
Olsen, Jr D. R. (2007). Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on user interface software and technology* (pp 251–258). ACM.
Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (2017). The readability of scientific texts is decreasing over time. bioRxiv 119370.
pypdf2. (2016). The PdfFileReader class. https://pythonhosted.org/PyPDF2/PdfFileReader.html. Accessed 15 Feb 2018.
Sawyer, A. G., Laran, J., & Xu, J. (2008). The readability of marketing journals: Are award-winning articles better written? *Journal of Marketing*, *72*(1), 108–117.
Wainer, J., Eckmann, M., & Rocha, A. (2015). Peer-selected best papers are they really that good? *PloS One*, *10*(3), e0118,446.
Wobbrock, J. O., & Kientz, J. A. (2016). Research contributions in human–computer interaction. *Interactions*, *23*(3), 38–44.
Young, A. (2016). Reflections on the games & play virtual committee meeting. http://sigchi.tumblr.com/post/143440663545/reflections-on-the-games-play-virtual-committee. Accessed 15 Feb 2018.
Zhang, Y., & Hornof, A. J. (2014). Understanding multitasking through parallelized strategy exploration and individualized cognitive modeling. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3885–3894). ACM.
Zimmerman, J. L. (1989). Improving a manuscript's readability and likelihood of publication. *Issues in Accounting Education*, *4*(2), 458–466.