An Analysis of IHC and HCII Publication Titles: Revealing and Comparing the Topics of Interest of their Communities

Franklin M. da C. Lima ⓘ **[Federal University of Rio Grande do Norte | franklinmatheus@ufrn.edu.br]**
Gabriel A. M. Vasiljevic ⓘ **[Federal University of Rio Grande do Norte | gabrielvasiljevic@ppgsc.ufrn.br]**
Leonardo Cunha de Miranda ⓘ **[Federal University of Rio Grande do Norte | leonardo@dimap.ufrn.br]**
M. Cecília C. Baranauskas ⓘ **[University of Campinas | cecilia@ic.unicamp.br]**

Abstract

Analyzing how the conferences of a given research field are evolving contributes to the academic community in that the researchers can better situate their research towards the advancement of knowledge in their area of expertise. Thus, in this work we present the results of a correlation analysis performed within and between-conferences of the field of Human-Computer Interaction, using data from the conference on Human-Computer Interaction International (HCII) and from the Brazilian Symposium on Human Factors in Computing Systems (IHC). More than 209 thousand words from the titles of over 18 thousand publications from both conferences were analyzed in total, using different quantitative, qualitative and visualization methods, including statistical tests. The analysis of words from the tiles of publications from both conferences and the comparison of the ranking of these words indicate, amongst other results, that there is a significant difference in relation to the main and most covered topics for each one of these conferences.

Keywords: *HCI, scientific community, papers, articles, titles, co-word.*

1 Introduction

The field of Human-Computer Interaction (HCI) is quite diverse in terms of conferences, having several scientific events worldwide. According to Gasparini et al. (2017), the 29 most proficient Brazilian researchers of the field, i.e., those with more publications in the Brazilian Symposium on Human Factors in Computing Systems (IHC), which is the main conference of the area of HCI in the Brazilian national scope, have also been publishing articles with considerable frequency in the HCI International (HCII). The HCII is, after the IHC, the conference in which the Brazilian researchers of the field of HCI most publish their works (Gasparini et al. 2017).

Since the first edition of the HCII, which occurred in 1984, the Conference has been presenting the most recent results of the field of HCI in each new edition. This conference has evolved over the years with specific characteristics that contributed to its current format. Since 2001, the HCII started to have affiliated conferences, i.e., it became the umbrella event of several satellite conferences of the field. Although each of these conferences have a specific focus, they all naturally follow the main theme of Human-Computer Interaction. Another important factor was that, since 2011, the conference started to provide a specific segment for the submission of extended abstracts and posters. Its affiliated conferences have also evolved over the years, but distinctly: some have expanded, others had their names changed, and some others ceased to exist. Currently, the HCII has 18 affiliated conferences.

Considering the importance of conferences to the dissemination of new knowledge and research results for the area of Computing, analyzing how the conferences of a subarea of Computing are evolving contributes for the researchers to have a clearer view of how their works relate to the state of the art and, thus, have better conditions to contribute to the advancement of knowledge in the field. The investigation of how a conference evolves along the years can be done with different methods and from different perspectives, depending on the objectives of the study. One possible way is to carry analyses based on the words the authors select for the titles of the publications. This type of study has the potential to produce a wide view of knowledge about the conferences, since the analysis is performed, in general, with data from several editions. The analysis of titles from conference publications has been an investigation strategy employed in various literature studies, such as in the works of Buchdid and Baranauskas (2014), Lima et al. (2018) and Liu et al. (2014).

Considering the natural importance of both the IHC and the HCII to the Brazilian scientific community of the field, the objective of this work is to analyze the propensities of the international community and compare them to those of the Brazilian community. The study presented in this work was performed through correlation analyses of the titles of publications from HCII and from IHC, with different quantitative, qualitative, and data visualization approaches. In addition, we also performed statistical tests to compare the individual ranking of words that appeared in the titles of publications in both conferences.

This article is organized as follows: Section 2 presents the related works; Section 3 describes the method employed in this study; Section 4 presents the results; Section 5 discusses the results; and Section 6 concludes the paper.

2   Related work

In this last decade, researchers from the field of HCI have been investigating how their scientific communities evolved through the years in various manners, e.g., through the application of questionnaires/surveys and the identification of co-authorship or scientific collaboration networks.

The literature works in the context of the Brazilian community were conducted with several different methods, ranging from the application of questionnaires/surveys for the community to analysis of the IHC publications from different perspectives (e.g., the GranDIHC-BR and its relation to the publications of the symposium). Still in the context of the Brazilian scenario, other similar studies were also done by scientific communities of other Computing areas. In the international context, this interest is the same, to investigate the evolution of scientific communities in area field from the perspective of a specific country or geographic region. These works are the most related to this work and are, therefore, described and cited in this section.

A complete analysis of the publications of the IHC is presented in Barbosa et al. (2017), including an analysis of the keywords from the publications. Nevertheless, the results were not compared with any other international conference of the field, making the context of the study focused exclusively on the scenario of the Brazilian community. In Buchdid and Baranauskas (2012) the IHC is analyzed with other conferences of the field, i.e., the Conference on Human Factors in Computing Systems (CHI) and the IFIP TC13 Conference on Human Computer Interaction (INTERACT), considering a publication period of five years, from 2007 to 2011. The analysis of the titles of the publications was carried out for these three conferences. However, the word analysis considered only one "dimension" of their relations, without considering the correlations between them.

Lima et al. (2018) presents an analysis of the words from the titles of IHC publications since its first edition, which occurred in 1998, and compared the conference to the CHI, also starting from its first edition, which occurred in 1981. The word analysis from the publications of both conferences, although using various methods of data visualization (i.e., tag cloud, heatmap and radviz), did not explore the correlation between the words from the titles.

Following the studies focused on analyzing the Brazilian symposium of the HCI field, Silva et al. (2018) sought to understand the impacts that the establishment of the Five Grand Research Challenges for the field of HCI on the publications of IHC. These challenges were established and presented by specialists from the Brazilian scientific community of the field, and are known as the I GranDIHC-BR (Baranauskas et al. 2015). In the study, the authors analyzed ten editions of the conference for the full papers that involved the theme of privacy, which is related to the challenge #4, and sought to show how the focus on this topic changed after the establishment of the I GranDIHC-BR. Following a similar line of study, Bueno et al. (2016), four years after the definition of the five grand challenges, sought to understand whether the works presented at the IHC during the years of 2013, 2014 and 2015 progressed according to

these challenges. For this purpose, 163 full and short papers were analyzed and correlated to the challenges, in order to show whether the researches were, in fact, following the proposals established in the I GranDIHC-BR. The research developed in Granatto et al. (2016) used a different approach, but also focused on analyzing studies that were published at IHC based on the I GranDIHC-BR. The authors conducted a literature review to analyze works published in the IHC that were related to digital accessibility. This analysis was aimed at understanding how the proposed challenges influenced the research focus on this theme, since the second challenge is specifically focused on accessibility.

In the work of Santana et al. (2017), the authors followed a similar line, and analyzed how the Brazilian community progressed according to the Challenge #1 (future, smart cities and sustainability) of the I GranDIHC-BR. The study addresses aspects related to this challenge in the IHC, indicating which obstacles may exist for the community to conduct research in each one of these aspects. Finally, the authors highlight possible directions that the community may follow to intensify research in this challenge.

In Coelho et al. (2017), the IHC was analyzed with other Brazilian events from the area of Computing, i.e., the Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), the Seminário Integrado de Software e Hardware (SEMISH), the Simpósio Brasileiro de Informática na Educação (SBIE) and the Simpósio Brasileiro de Engenharia de Software (SBES). In this work, the authors analyzed the publications of these events from the perspective of the Challenge #4 from the Brazilian Computer Society (SBC) (Baranauskas and Souza 2006).

In addition to the IHC, Brazilian researchers from other areas of Computing that have specific local conferences also try to understand their communities, which is the case of the Simpósio Brasileiro de Sistemas Colaborativos (SBSC). In the study presented in Steinmacher et al. (2013), for example, the SBSC was analyzed and issues related to the themes of the publications were presented and discussed, in addition to the co-authoring networks of these works. In this study, the authors sought to answer questions related to the topics covered in the analyzed works, while also indicating the changes that occurred in these topics over time and in general.

In the international context, Padilla et al. (2014) studied five editions of the CHI aiming at understanding what are the main topics covered by this conference. A new visualization method was proposed in this work, called Trend Map, which is capable of demonstrating which topics are stable, trending or in decline. These trends are indicated with clouds of words for each identified topic.

As in Brazil, researchers seek to analyze regional conferences from other parts of the world. In the work of Mubin et al. (2017), for example, the Australian Conference on Human-Computer Interaction (OzCHI) was analyzed from various perspectives. The authors sought to understand issues related to the acceptance rate and number of submissions per edition of the conference, as well as who were the authors of the published studies and their affiliations. The most popular topics and the most cited

works brought important information about the Australian conference. In addition, the OzCHI was also compared to other regional conference, the Nordic Conference on Human-Computer Interaction (NordiCHI). Another study regarding regional conferences is described in Gupta (2015), in which the analyzed conference is the India Human Computer Interaction (IndiaHCI).

Regarding the related studies, our work conducts a different analysis, in terms of both the covered conferences and the objectives of the investigation. In addition, different focuses and strategies of analysis were employed, as the titles of publications from the HCII and from the IHC were analyzed through the perspective of words and terms, with one or more dimensions in their relations, i.e., ranking, correlation of pairs of words, sequences of words, and terms. The study used statistical methods and graphical visualizations that correlated words and terms in matrices, graphs and *n*-grams networks, according to the method described in the next section.

3 Method

For this work, considering its objectives as defined in Section 1, the following method was established, which is composed of five sequential steps of data manipulation: (i) selection of data sources; (ii) data collection; (iii) pre-processing; (iv) processing and visualization; and (v) statistical analysis.

3.1 Selection of data sources

The first step in this study was to investigate from which sources the titles of the publications from both the HCII and the IHC would be collected. Naturally, as a first option of data source, we decided to verify availability in the digital libraries of the conferences' respective publishers.

In relation to the HCII, it is worth mentioning that this conference, from its first edition to the most recent one, had different publishers, i.e., Elsevier (1984, 1987, 1989, 1991, 1993, 1995, 1997), Lawrence Erlbaum Associates (1999, 2001, 2003, 2005) and Springer (2007, 2009, 2011, 2013-2020). No information was found in the digital libraries of the Elsevier and the Lawrence Erlbaum Associates, that is, the data from 11 editions of the HCII were not available for access through the Internet. The only publisher of the HCII which maintains all its conference proceedings is Springer, with a high number of published works.

In relation to the IHC, its publisher is the Association for Computing Machinery (ACM), but not all the conference proceedings are available in its digital library (ACM DL). However, as the IHC is promoted by the Human-Computer Interaction Special Committee (CEIHC) [1] of SBC, that Committee keeps a website with various information of interest for the Brazilian community, including all the proceedings of the IHC since 1998, when its first edition occurred.

The Digital Bibliography & Library Project (DBLP) [2], which is a digital bibliographic repository that aggregates proceedings of several conferences in the area of Computing, was also investigated as a possible data source. The DBLP was checked to verify the possibility of employing a single source for the data collection of both conferences. It was noticed that the DBLP had data of all editions of HCII published by Springer, in addition to some others (five from a total of 11 editions) that were published by Elsevier and by Lawrence Erlbaum Associates; however, not all of them were complete. In relation to the IHC, it was also noted that not all data from this conference was available in the DBLP, since there was no information of the proceedings published before 2006.

3.2 Data collection

Based on the data sources that were selected to gather the titles of publications from HCII and IHC, i.e., the DBLP and the CEIHC/SBC portal, respectively, the next step was to proceed to the data collection itself.

The process of data collection using web scraping was executed as follows: (i) investigation of the source code of the data sources' HTML pages, aiming at verifying where and how the titles were stored in those pages; (ii) manual gathering of the URL of all pages that contained the titles; (iii) development of scripts using Node.js [3], which accessed the URL links obtained in the previous step to automatically extract the titles based on the information gathered in step (i); and (iv) creation of a local structured database with all the collected titles, including additional information such as conference edition and location from which the article was published. Before the end of this step, with the database already populated, the titles of a random sample set of collected data were checked to verify the completeness and correctness of the data collection.

3.3 Data pre-processing

With the collection and structuring of all the titles from the publications, the next step was to pre-process the titles, which was performed in two different sub steps. All the pre-processing and the generation of results were performed using the Jupyter Notebook [4] and the Python programming language.

The first stage of the pre-processing involved clearing the data through the removal of the following groups of data from the titles: (i) stopwords, i.e., words that do not carry significant meaning (e.g., connectives, possessive pronouns and personal pronouns); (ii) punctuation characters (e.g., commas, dashes and colons); and (iii) numerical values, that is, only numbers (e.g., "10" e "2020"), as numbers conjugated with letters were considered (e.g., "2d" and "3dtv"). As a result of this step, the titles were left only with words that contain any semantic meaning and are of interest for this study. We will further reference them as "significant words". The Language Toolkit (NLTK) [5] was employed for

---

[1] comissoes.sbc.org.br/ce-ihc/.
[2] dblp.org.
[3] nodejs.org.

[4] jupyter.org.
[5] nltk.org.

this procedure, which is a Python library that processes natural human language data.

The NLTK returns a series of stopwords from a given input language. In the specific case of the HCII, the given language was English, as all its publications' titles are in this language. Some examples of the dozens of stopwords that the library returns for this language are "I", "me", "my", "myself", "we", "our", "ourselves", "you" and "you're". For the case of the IHC, in which titles in Portuguese are also present, stopwords in this language also had to be removed. Some examples of stopwords that the NLTK returns for this language are "de", "que", "nosso", "para", "não", "quando", "também", "qual" and "você". In addition to stopwords, the NLTK also provides a set of punctuation characters that were also removed from the titles, such as "!", ":", "&" and "?".

It was initially planned to substitute every occurrence of stopwords for empty strings; however, we noticed that this would not be enough for every case, as there were instances of words that were separated by hyphens and should be considered individually, or words with apostrophes, such as "user's", which should be considered as "user" given the objective of acquiring only the significative words from the titles. To avoid situations like these, punctuation characters were removed first, being replaced with white spaces. This resulted in cases of many white spaces in sequence, which were also removed. Then, with all words separated by only one white space and without any punctuation character, all stopwords and other unnecessary terms, such as numerical values, were removed.

The last step of the pre-processing was performed after this procedure of removing stopwords, punctuation characters and numerical values. In this step, each title was transformed into a sequence of tokens, which are the words that carry more significant semantic values. Organizing the titles in tokens was a fundamental step for the pre-processing of data, since the following steps were simplified by this arrangement.

3.4 Data processing and visualization

Based on the results obtained from the pre-processing stage, with the titles organized in tokens (words that carry semantic meanings), the next step involved generating results that would assist in the visualization of correlations between these tokens of words. This correlation is defined by the proximity of the tokens in the titles, i.e., for each token T in a title, T is correlated with its immediate neighbors $V1$ and $V2$. Notice that tokens positioned at the beginning or at the end of the corresponding title have only one neighbor.

To produce these results, the NLTK was used again to generate n-grams of the titles. n-grams are, in summary and in the context of this work, sequences of n tokens composed in each title. For example, considering the title of a random publication from HCII which have the token sequence ["challenges", "user", "centered", "smart", "environments"], its 3-grams would be ["challenges", "user", "centered"], ["user", "centered", "smart"] and ["centered", "smart", "environments"].

Based on these n-grams, three visualizations were planned to illustrate the correlation of tokens of words: the word correlation matrix, the term correlation graph and the n-grams correlation network. The methods employed to generate each of these visualizations are described as follows.

3.4.1 Word correlation matrix

The word correlation matrix is a heatmap that indicates and quantifies the correlation of pairs of words in the same title through the intensity of the colors in its cells. As complementary information, the number of times that each word appears in the titles is also presented. These words may appear in different positions in the title, and not necessarily in sequence. For example, in a given title, a word $W1$ could appear immediately before some other word $W2$ and, in another title, these same words could be positioned in an inverse order and not necessarily in sequence (e.g., in some other title, the word $W2$ could appear before a word $W3$, and both of them positioned before word $W1$). In this example, therefore, this pair of words ($W1 \leftrightarrow W2$) have two correlation occurrences, as it does not depend on the sequence and/or the order in which they appear.

The word correlation matrix is a square matrix of order N, being N the number of significative words that, when sorted in descending order in relation to their absolute frequency (fi), forms the ranking of these N words (#P, being P its position in the ranking). The value of N is defined empirically by considering the percentage of the cumulative relative frequency ($Fri\%$) of the sample, as it allows the analysis of the N most recurring words which, in turn, also potentially represent the most recurring topics from the conferences. Since this matrix is symmetric, all elements above the main diagonal were removed to avoid duplicates. Therefore, this matrix presents the following number of possible correlations between the N words (Equation 1):

$$\frac{N * (N - 1)}{2} + N \qquad (1)$$

For the production of the correlation matrix, as previously described, the 2-grams of the titles were initially generated with the NLTK and then applied to a library of data visualization, the Seaborn[1], in order to generate the heatmap itself. After the heatmap was generated, the scripts were altered for the inclusion of complementary information, such as ranking, total number of occurrences and percentage of occurrences within the matrix, in addition to minor esthetic adjustments for the final version of the correlation matrix, which is presented in this work.

The publications of the IHC may have titles in three different languages, i.e., English, Portuguese and Spanish. For the elaboration of the word correlation matrix of this conference, the titles were kept in their original language, since any translation process could generate noise in the results. Thus, words from different languages may appear in the visualization of the IHC word correlation matrix.

---

[1] seaborn.pydata.org.

3.4.2 Term correlation graph

The term correlation graph was created to complement the analysis of the word correlation matrix. The VOSViewer[1] software was used for this task, since it possesses specialized algorithms for the identification of terms. In this work, terms can be defined as simple or composite words (which are formed by two or more words) and are presented as vertices in a graph. In this visualization, vertices formed by composite words have a stronger relation between the words that compose them, in comparison to the connection that the simple words that compose them would have if they were connected on their own. The vertices are also organized by clusters in the graph through specialized clustering algorithms, as a way to better group vertices with greater similarities to each other.

VOSViewer is a free software tool developed to facilitate bibliometric analysis, being focused on visualizations of bibliometric mappings (Van Eck and Waltman 2010). Several studies in the literature employed this tool, either to highlight research topics in a particular area (e.g., Aghimien et al. 2019) or to investigate citation networks (e.g., Leydesdorff et al. 2013). In the context of the present work, the VOSViewer was employed to generate term correlation graphs, in which the relations between the terms are represented by connections (edges).

The tool has a certain learning curve for its best use and, based on the textual data it receives, is able to identify the terms that should be presented in the resulting graph visualization. The software employs natural language processing algorithms to identify these terms and, for these algorithms to function correctly, it is necessary that all textual data are in the English language (Van Eck and Waltman 2013). Thus, since the VOSViewer is capable of processing only English textual data, it is important to highlight that, in the context of the IHC analysis, the titles had to be separated by language so that only those in English were processed to generate the term correlation graph.

The process for generating the term correlation graph can be defined as a sequence of four steps: (i) to define the input data file that contains the titles of the publications; (ii) to define the method for counting the terms; (iii) to define the limit for the number of terms (based on the minimum number of occurrences); and (iv) to define the number of terms that would appear in the visualization. In step (i), the data file should contain all titles that will be analyzed. The counting method defined in step (ii) was the counting by number of occurrences of the terms. In step (iii), the number of terms were defined based on all titles on the data file, since the value of N (number of terms) is defined empirically considering the percentages of the cumulative relative frequencies (*Fri%*) of the sample. This allows the analysis of the N most recurring terms, which are also the terms related to the most recurring topics from each conference. In this sense, the minimum number of occurrences for each term (bottom) defined in step (ii) was chosen as follows: given that a *Fri%* of the most recurring words of the conference equals to N, the bottom must be defined in such way that at least N terms identified by the tool present a number of occurrences equal or higher than bottom. Finally, in the last step (iv), it was defined that all terms that were selected based on the filtering performed in step (iii) would appear in the visualization by default.

3.4.3 *n*-grams correlation network

The n-grams correlation network is a visualization that presents a list for each set of *n*-grams of different sizes, in which each set is ranked based on the number of occurrences of its instances. As additional information, the *n*-grams were connected to the (n+1)-grams in which they appear, that is, if a pair of words that appears in the set of 2-grams also appears in any 3-grams, there will be a connection between this 2-gram and the 3-gram, and so on.

The first step in generating the *n*-grams correlation network is to find what is the highest possible n to generate *n*-grams that does not have very similar or low number of occurrences. Starting from the 2-grams, new sets of *n*-grams were generated individually by incrementing the value of n by 1 successively, until the number of instances of a given set of *n*-grams were not repetitive. This procedure was performed separately for each conference.

The next step was to define how many and which *n*-grams would be selected to appear in the network visualization. A similar approach to the one employed for the correlation matrix was used, i.e., the *n*-grams were classified by descending order of occurrences and a cutoff point was determined empirically depending on the sample. This cutoff point was defined by a process divided into two steps. First, given that the number of *n*-grams was relatively high, all those which had a number of occurrences equal to 1 were discarded. The first *n*-grams from each ranking that had a cumulative frequency of the number of occurrences equal to a percentage of the sample size were selected. The second and last step of defining the cutoff point was performed after the *n*-grams were selected. In this step, the cutoff point would increase or decrease in the ranking of the respective *n*-grams if it was positioned in the middle of a sequence of repeated number of occurrences. The following rules were defined for this step: (i) if the sequences of repeated numbers are equal to two occurrences by *n*-gram, the cutoff point increases to the last element in the ranking that has a number of occurrences equal or higher than 2; (ii) if the numbers in sequence are higher than two occurrences, the cutoff point decreases to the last element of the ranking with this same number of occurrences.

After this second step of defining the cutoff points, the ordered sets of n-grams and the n-grams correlation network were created. The sets were arranged so that the connections from a given *n*-gram to the (n+1)-gram in which it appears were clearly visible by a line that links these two *n*-grams. For this purpose, all these connections from one set of *n*-grams to the set of (n+1)-grams had to be posteriorly added to the image using a vector image editor. The entire process

---

[1] vosviewer.com.

of organization and preparation of the images was done manually for each individual connection.

As with the word correlation matrix, the titles of the publications were kept in their original language in the *n*-grams correlation network of the IHC, which may result in *n*-grams of three different languages (i.e., English, Portuguese and Spanish) in the corresponding visualization.

3.5 Statistical analysis

As fifth and last step of the method, we performed statistical tests aimed at complementing the previous analysis, and to verify the differences between the topics of the IHC in comparison to those of HCII in a more objective manner. For this comparative analysis, we performed agreement and concordance tests between the ranking of words from both conferences. The words in Portuguese from IHC publications were translated and their total number of instances were added to their English counterparts, thus allowing the comparison between all words that appear in both conferences.

For the analysis of concordance, we employed an Inter-Rater Reliability (IRR) metric, which considers the ranking of words from each individual conference as a metric of agreement between them. The measure of IRR that we used to infer this concordance was the Intraclass Correlation Coefficient (ICC), which is a reliability index widely employed in both within- and between-raters agreement analysis (Koo and Li 2016). The coefficient was calculated as a two-way mixed model and for absolute agreement.

The Kendall's tau-b (τ_b) non-parametric correlation coefficient was employed for the correlation analysis. This coefficient measures the proportion of concordant pairs relative to discordant pairs, i.e., the proportion in which the rankings of both conferences correlate to each other as the same words are classified in similar positions. All statistical tests were performed using the IBM SPSS Statistics [1] software.

4 Results

This section presents the results of this work based on the method described in the previous section. The results were first described individually for each conference, with the analysis of the correlations of (i) words, (ii) terms, and (iii) sequences of words in the titles of their publications. The results from the comparison between the rankings of words of both conferences is presented after these analyses.

4.1 HCII

A total of 17.454 articles were identified for this conference, which were published in its 11 editions that occurred between 2007 and 2020 ($M = 1.586,73$; $SD = 179,60$). The set of all the words that compose the titles of these publications consists of 201.379 words.

4.1.1 Analysis of words from HCII

From the universe of words coming from the titles of HCII's publications (201.379 words), and not considering the stopwords since they do not carry value in the context of this analysis, the significative words from this conference were identified, being **N1 the set of all different significative words from HCII**, i.e., 12.448 words.

Figure 1 represents the distribution of N1 by its absolute frequency (*fi*) in descending order. The percentages of cumulative relative frequency (Fri%) are 12,95% (n = 15), 25,24% (n = 52), 50,03% (n = 276) and 75,00% (n = 1.116). Thus, the most recurring significative words from HCII are concentrated in the first positions of the histogram, being **n1 the set of the first *n* words with highest *fi* from N1 and with a *Fri%* of approximately one fourth of the total**.
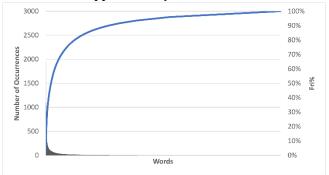

Figure 1. Histogram of the N1 (12.448) words from HCII.

The correlation analysis of the words from HCII was performed with the creation of the correlation matrix of its words with 52 dimensions (Figure 2), since the 52 most recurring words amongst the 12.448 (N1) significant words from HCII concentrate, alone, approximately 25% of the total of occurrences. Being so, this matrix presents the 1.378 possible correlations between the 52 most recurring words from the HCII, and each of them have 351 or more occurrences.

The words are presented as labels in the rows to the left of the matrix, preceded by its position in the ranking. The words are presented from top to bottom, going from the word with highest *fi* (#1) to the word with lowest *fi* (#N). The labels also show the total number of occurrences of each word and the percentage of its occurrences within the matrix. The labels of the columns (below the matrix) also present words preceded by their positions in the ranking, for better visualization of the correlation between pairs of words considering the axes of the matrix. The cells of the matrix (#L,C, being *L* the row position and *C* the column position) are painted with a color that varies according to the frequency of its occurrence, from a lighter background (zero occurrences) to a darker background (highest value of occurrences), and represented by the color scale - Color MAPping (CMAP) – at the right side of the matrix.

Each axis of this matrix contains the ranking of words from HCII, which presents the following words with more than 1.000 occurrences (in descending order): "design" (#1; 2.657), "based" (#2; 1.946), "user" (#3; 1.666), "using" (#4; 1.483), "system" (#5; 1.465), "study" (#6; 1.371), "human" (#7; 1.092) and "interaction" (#8; 1.087). The analysis of this

---

ranking helps to reveal the most explored topics of this conference. However, analyzing words in an isolated manner could not be enough to identify the topics of most interest for HCII, as words such as "based" (#2; 1.946), "using" (#4; 1.483), "use" (#42; 391) and "new" (#49; 354), lack an appropriate complement for understanding more precisely

what meaning or related topic these words could carry when complemented with some other word. Therefore, analyzing the combined use of the most recurrent words present in the publication titles of HCII can bring to light new results on the most discussed topics at this conference.

Figure 2. Correlation matrix of the n1 (52) most recurring words from HCII.

In this sense, the analysis of correlation between two related words in this matrix can contribute to reveal new subjects covered in this conference. The 17 correlations between pairs of words in this matrix that have more than 100 occurrences are (in descending order): "user ↔ experience" (#3,18; #18,3; 349), "virtual ↔ reality" (#10,19; #19,10; 333), "study ↔ case" (#6,29; #29,6; 330), "augmented ↔ reality" (#46,19; 267), "interface ↔ user" (#14,3; 227), "human ↔ computer" (#7,26; #26,7; 200), "interfaces ↔ user" (#38,3; 159), "interaction ↔ computer" (#8,26; 154), "design ↔ interaction" (#1,8; 147), "evaluation ↔ usability" (#11,17; #17,11; 126) e "system ↔ support" (#5,33; #33,5; 117). Some of these correlations of words could be considered as expected, such as "computer ↔ human" and "computer ↔ interaction", as they are related to the acronym of the field of IHC and are, therefore, guiding keywords for the topics of interest of the area. Thus, the

analysis of pairs of words from this conference reveals interesting results, such as "user experience" being a prominent topic in the studies. It is also clear that many works presented studies of interactions based on technologies of virtual reality or augmented reality, and regarding methodologies, the results suggest that most studies are predominantly performed as case studies, and the most explored evaluation is that of usability.

In addition to the correlations of words analyzed previously (with a 100 or more occurrences), other correlations between pairs of words in this matrix also contribute to the identification of other subjects of interest in this conference, even with a smaller number of occurrences. The 35 correlations between pairs of words with less than 100 occurrences are the following (in descending order): "using ↔ system" (#4,5; 83), "service ↔ design" (#51,1; 81), "environments ↔ virtual" (#52,10; 77), "based ↔

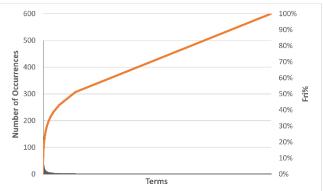system" (#2,5; 74), "web ↔ based" (#36,2; 73), "mobile ↔ application" (#12,34; #34,12; 65), "model ↔ based" (#21,2; 63), "environment ↔ virtual" (#39,10; 60), "research ↔ design" (#22,1; 55), "approach ↔ design" (#28,1; 51), "learning ↔ based" (#9,2; 50), "interactive ↔ design" (#23,1; 47), "information ↔ system" (#15,5; 46), "training ↔ system" (#48,5; 41), "systems ↔ information" (#24,15; 40), "technology ↔ information" (#30,15; 36), "development ↔ design" (#20,1; 34), "method ↔ design" (#32,1; 34), "performance ↔ human" (#37,7; 34), "game ↔ design" (#50,1; 32), "social ↔ interaction" (#16,8; 31), "digital ↔ human" (#27,7; 31), "smart ↔ environments" (#43,52; 27), "communication ↔ system" (#45,5; 26), "online ↔ social" (#41,16; 24), "analysis ↔ usability" (#13,17; 23), "new ↔ approach" (#49,28; 20), "data ↔ analysis" (#25,13; 19), "use ↔ social" (#42,16; 19), "cognitive ↔ based" (#47,2; 17), "visual ↔ design" (#35,1; 14), "people ↔ visual" (#44,35; 13), "designing ↔ interactive" (#40,23; 12) and "users ↔ based" (#31,2; 9). These additional correlations help to reveal more topics of interest of this conference, being clear the strong interest in studying social interactions, in addition to the analysis of interactions in intelligent and virtual environments. It is also clear that there many studies that explore data analysis and human performance, and that approach several issues related to the applications of mobile devices and information systems, in addition to the interest in design in a variety of fields, such as games. It can still be noticed, that the combined use "based" with many other words helps to clarify what the authors emphasize in regard to "based", i.e., that many works relate to being based on web, or based on models, or based on learning, for example. These correlations also explain works that present new approaches, design methods or design approaches.

The different word correlation analyzes, as presented above and which are most easily viewed in the two dimensions of the correlation matrix in Figure 2, revealed diversified and recurring topics of interest to HCII.

4.1.2 HCII terms analysis

The terms of the HCII were identified from the total of 17.454 titles of its publications, being **N2 the set of all different terms of the HCII**, i.e., 36.950 terms. The histogram in Figure 3 presents the distribution of N2 based on the absolute frequency (fi) of its terms in descending order. The percentages of the cumulative relative frequency ($Fri\%$) are 12,51% (n = 58), 25,01% (n = 355), 50,00% (n = 4.879) and 75,00% (n = 20.731). The most recurring terms of the HCII are also concentrated in the first positions, being **n2 the set of the first n terms with highest fi from N2 and with $Fri\%$ equal to approximately one eighth of the total**. It is worth mentioning that the 59th term with highest fi have exactly the same fi of the 58th term; therefore, both were considered in this analysis.

The term correlation analysis of the HCII was performed from the creation of the term correlation graph of this conference, with a total of 59 vertices (Figure 4), since the 59 (n2) most recurring terms concentrate over 12,5% of all occurrences from the total of 36.950 (N2) terms. Thus, this

graph shows the 1.343 connections (edges) between the 59 vertices that represent the most recurring terms from HCII, and each have 61 or more occurrences. The size of each vertex is proportional to the number of occurrences of its corresponding term, and the vertices are organized in clusters to better highlight the proximity between related terms.

Figure 3. Histogram of the N2 (36.950) terms from HCII.

In the term correlation graph from HCII, the terms were divided into four clusters (the different colors of the vertices), described next in descending order of occurrences. The cluster C1, in red, is composed of 27 terms: "study" (1.016), "effect" (613), "research" (472), "technology" (447), "experience" (381), "performance" (331), "case study" (328), "use" (322), "information" (240), "perspective" (233), "communication" (229), "impact" (208), "influence" (195), "review" (189), "perception" (170), "comparison" (157), "older adults" (157), "context" (154), "difference" (153), "practice" (143), "role" (131), "relationship" (114), "china" (112), "human computer interaction" (108), "social media" (99), "hci" (83) and "example" (67). C2, in green, is composed by the following 16 terms: "analysis" (782), "evaluation" (644), "development" (586), "approach" (549), "model" (522), "application" (515), "framework" (329), "usability" (318), "tool" (266), "user interface" (233), "assessment" (201), "visualization" (194), "case" (148), "accessibility" (142), "web" (129) and "implementation" (106). C3, in blue, is composed by 9 terms: "system" (1.488), "person" (375), "children" (183), "virtual reality" (163), "elderly" (141), "proposal" (135), "augmented reality" (125), "usability evaluation" (100) and "gamification" (79). C4, in yellow, consisted of 7 terms: "interaction" (657), "interface" (533), "user" (286), "user experience" (250), "investigation" (131), "mobile device" (110) and "designing" (105).

Notice that there are vertices of terms that represent composite words consisting of two or three words (in descending order of occurrence): "case study" (328), "user experience" (250), "user interface" (233), "virtual reality" (163), "older adults" (157), "augmented reality" (125), "mobile device" (110), "human computer interaction" (108), "usability evaluation" (100), "social media" (99). This demonstrates the strong correlation that these words have with each other, despite the fact that most of these correlations have already been identified through the word correlation matrix. Note, however, that terms that are related to other topics of interest of this conference appeared in this

analysis and did not appear in the word correlation matrix. Thus, this term correlation graph complements the results that were obtained previously in the word correlation matrix.

Despite the small difference between the amount of terms from the correlation graph (59) and the correlation matrix (52), 30 new terms that did not appear as words in the correlation matrix emerged in the graph, which are the following, presented in descending order of occurrence: "effect" (613), "person" (375), "framework" (329), "tool" (266), "perspective" (233), "impact" (208), "assessment" (201), "influence" (195), "visualization" (194), "review"

(189), "children" (183), "perception" (170), "comparison" (157), "older adult" (157), "context" (154), "difference" (153), "practice" (143), "accessibility" (142), "elderly" (141), "proposal" (135), "investigation" (131), "role" (131), "relationship" (114), "china" (112), "device" de "mobile device" (110), "implementation" (106), "media" de "social media" (99), "hci" (83), "gamification" (79) and "example" (67). The emergence of these terms in this visualization shows other topics of interest for the community, such as the conducting of reviews, studies involving comparisons, the concern with aspects related to accessibility, an older target audience, and gamification issues in interactive systems.
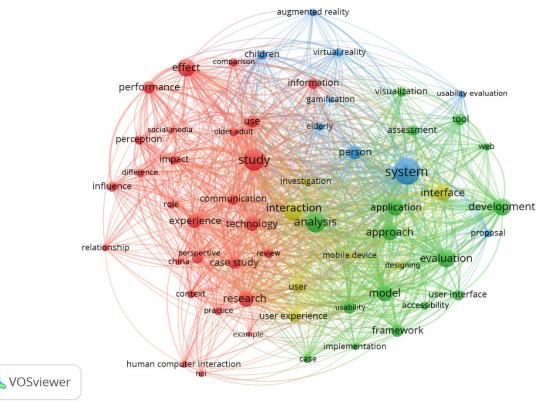

Figure 4. Correlation graph of the n2 (59) most recurring terms of HCII.

4.1.3 Analysis of sequence of words from HCII

The *n*-grams for this conference were generated from a total of 17.454 titles from its publications, being **N3 the set of all different instances of a given *n*-gram from HCII**. The N3 of the different *n*-grams from HCII are: 2-gram (12.917), 3-gram (3.392), 4-gram (761), 5-gram (288), 6-gram (153) and 7-gram (86). The histograms in Figure 5 show the distribution of each of these N3 by their absolute frequency in descending order, while Table 1 presents the different percentages of cumulative relative frequencies (*Fri%*). These histograms are less concentrated than those of the significative words from HCII (Figure 1), and also show that as the number of sequences of words (*n*) that compose a given *n*-gram increases, the distribution tends to become more homogenous. For this analysis, **n3 is the set of all first *n* instances of a given *n*-gram with highest fi from N3 and with *Fri%* equal to approximately one eighth o the total**.

It is worth highlighting that the sample size (N3) of each different *n*-gram from HCII varies and, therefore, n3 also

varies, according to the rules defined in Section 3 (more details in Section 3.4.3). For analysis purposes, it is also worth mentioning that the value of n3 was chosen so that the *Fri%* is in the order of one eighth of the total, and that n ≥ 3.

The analysis of sequence of words from HCII was performed through the creation of the *n*-grams correlation network for this conference (Figure 6). Considering the data from HCII, and considering the method described in Section 3, the network connects six different ordered sets of *n*-grams, i.e., 2-gram, 3-gram, ..., 6-gram, 7-gram.

Considering the 12.917 (N3) possible 2-grams from HCII, the instance in the 73rd position of the ordered set of 2-grams has $fi = 40$. As the bigram must present all instances with this same fi, it was generated with instances up to the 74a position of this set (Figure 6a). As for the 3.392 (N3) possible 3-grams from HCII, the instance in the 64th position of the set of 3-grams has $fi = 9$. As the trigram must present all instances with this same fi, it was generated with instances up to the 72nd position of the set (Figure 6b).

Considering the 761 (N3) possible 4-grams from HCII, the instance in the 49ª position of the ordered set of 4-grams have fi = 3. As the quadgram must present all instances with this same fi, it was generated with instances up to the 99th position (Figure 6c). For the 288 (N3) possible 5-grams, the instance in the 22nd position of the set of 5-grams has fi = 2. As the pentgram must present all instances with $fi \geq 3$, it was generated with instances up to the 21st position (Figure 6d), as all instances starting from the 22nd 5-gram have $fi \leq 2$.

Table 1. N3 and n3 of the different *n*-grams from HCII.

n-gram	Fri%				N3	n3
	12,5%	25%	50%	75%		
2-gram	73 (12,56%)	365 (25,01%)	2.192 (50,00%)	6.497 (75,00%)	12.917	74
3-gram	64 (12,57%)	264 (25,00%)	1.060 (50,01%)	2.226 (75,01%)	3.392	72
4-gram	49 (12,52%)	130 (25,10%)	340 (50,03%)	551 (75,07%)	761	99
5-gram	22 (12,50%)	60 (25,00%)	136 (50,00%)	212 (75,00%)	288	21
6-gram	15 (12,66%)	35 (25,32%)	74 (50,00%)	114 (75,32%)	153	6
7-gram	9 (13,48%)	2 (25,84%)	42 (50,56%)	64 (75,28%)	86	3

Considering the 153 (N3) possible 6-grams, the instance in the 15th position of the ordered set of 6-grams has fi = 2. Since $fi \leq 2$ starting from the 7th position of this set, the hexagram was generated up to the 6th position, as it must present all instances with $fi \geq 3$ (Figure 6e). For the 86 (N3) possible 7-grams from HCII, the instance in the 9th position of the set of 7-grams has fi = 2. Since $fi \leq 2$ starting from the 4th position of this set, the heptagram was generated up to the 3rd position, since it must present all instances with $fi \geq 3$ (Figure 6f).

In the *n*-grams correlation network, it is possible to observe that, in many cases, a given *n*-gram comes from the set of (*n*-1)-grams and goes to the set of (*n*+1)-grams, which ends up creating a navigation sequence between three *n*-grams in the network. This may indicate how a subject still remains amongst the most recurring topics even though it is a specificity of a more general topic with high importance. The 3-gram "using – augmented – reality" (#9; 24) is one of such cases, which is a specificity of a more general topic indicated by the 2-gram "augmented – reality" (#3; 264) and that, in turn, appears in the 4-gram "using – augmented – reality – technology" (#37; 3), which indicates an even more specific topic.

It is also worth mentioning the *n*-grams that weren't derived from any of the (*n*-1)-grams and does not appear in any of the (*n*+1)-grams, since they may be related to specific subjects that are not related to more general ones, which are present in the (*n*-1)-grams, and are also not related to the more specialized ones in the set of (*n*+1)-grams. These *n*-grams appear, exclusively, in the sets of 3-grams and 4-grams, given that nothing precedes the 2-grams and for every 5-grams, 6-grams and 7-grams, there is a connection coming from a preceding *n*-gram. For the cases of such 3-grams, it is worth mentioning "adaptive – instructional – systems" (#16; 18), "nuclear – power – plants" (#33; 13), "digital – human – modeling" (#41; 12), "user – centred – design" (#47; 11) and "simulation – based – training" (#51; 10). The main 4-grams in such cases are "advanced – driver – assistance – systems" (#6; 6), "generalized – intelligent – framework – tutoring" (#10; 5), "global – public – inclusive – infrastructure" (#11; 5), "social – live – streaming – services" (#12; 5) and "eye – gaze – input – system" (#14; 5).
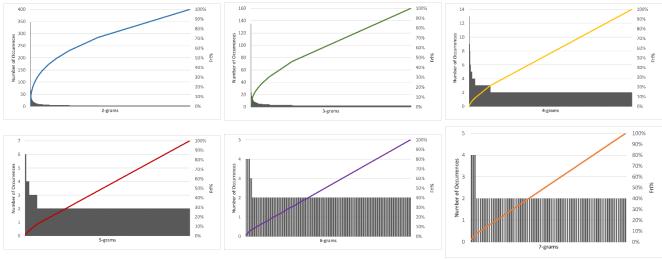

Figure 5. Histograms of the different *n*-grams from HCII: (a) bigram; (b) trigram; (c) quadgram; (d) pentgram; (e) hexagram; (f) heptagram.

Another phenomenon that can be observed in the network from HCII are *n*-grams that do not appear in the set of (*n*+1)-grams but do appear in the set of (*n*+2)-grams or any other posterior set. This may indicate more specialized topics of a general one that only gets to be among the main topics when it becomes even more specific. For example, the 2-gram "design – approach" (#66; 43) does not appear in any 3-grams, but is present in the 4-gram "user – centered – design – approach" (#13; 5).
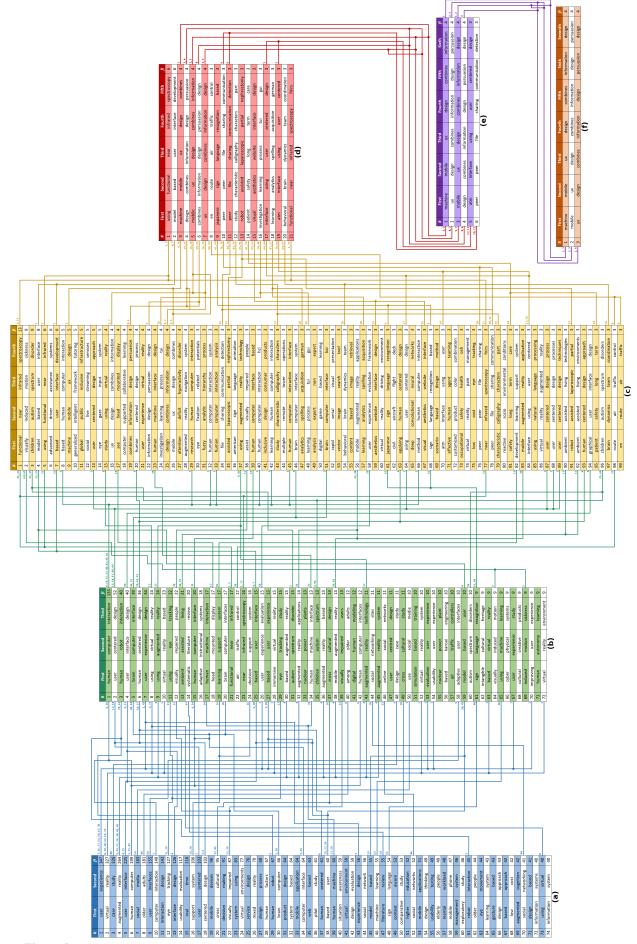
Figure 6. *n*-grams correlation network from HCII: (a) bigram; (b) trigram; (c) quadgram; (d) pentgram; (e) hexagram; and (f) heptagram.

There are connections between all six sets of *n*-grams of the *n*-grams correlation network from this conference. Identifying *n*-grams that are related to mode general subjects (lower *n*) that appear in several posterior *n*-grams (higher *n*), may reveal topics of high relevance that remain between the main subjects even when becoming more specific. This phenomenon forms a specie of chain of *n*-grams, starting from a lower *n* and gradually progressing to a higher one. Among the chains that are present in the network, two of them were identified as being the largest ones. They start from the set of 2-grams and progress to the set of 6-grams, forming sequences of five *n*-grams that differ only in their respective 2-grams. In one of these chains, the 2-gram is "user – centered" (#17; 103) and, in the other, it is "centered – design" (#18; 102). It is worth mentioning that the 2-grams from both of these chains converge into the same 3-gram. The other *n*-grams from both chains are the same: "user – centered – design" (#2; 52), "using – user – centered – design" (#64; 3), "interface – using – user – centered – design" (#17; 3) and "atm – interface – using – user – centered – design" (#5; 3).

The *n*-grams that converge to the same (*n*+1)-gram can also be highlighted, as they may indicate how two or more general topics also converge to a single more specific one. Considering 2-grams, for example, "human – computer" (#6; 199) and "computer – interaction" (#10; 199) converge to "human – computer – interaction" (#1; 135); "user – centered" (#17; 103) and "centered – design" (#18; 102) to "user – centered – design" (#2; 52); and "human – robot" (#29; 67) and "robot – interaction" (#61; 45) to "human – robot – interaction" (#3; 40).

Throughout the entire HCII *n*-gram network, it is possible to notice the presence of *n*-grams that do not appear in any other posterior set. This may indicate how general topics diverge into more specific ones that are not amongst the most recurring ones. One example for the case of 2-grams is "interaction – design" (#11; 142), which although is one of the most recurring 2-grams, does not appear in any posterior n-gram of the network, indicating that it spreads into more specialized topics with lower numbers of occurrences.

The analysis of the bigrams that does not appear in any of the (*n*+1)-grams shows a diversity in the topics. It is possible to observe that a large quantity of 2-grams appears in the 3-grams, i.e., from the 74 bigrams, 44,59% (33) of them appears in a 3-gram. From these 2-grams, "user – experience" and "virtual – reality" stand out as they appear the most in the 3-grams (seven and six times, respectively). For the case of "user – experience", the associated 3-grams are "user – experience – design", "user – experience – evaluation", "based – user – experience", "evaluation – user – experience", "usability – user – experience", "user – experience – study" e "improving – user – experience". For "virtual – reality", the associated 3-grams are "using – virtual – reality", "virtual – reality – based", "immersive – virtual – reality", "virtual – reality – system", "based – virtual – reality" e "virtual – reality – environment".

From these 3-grams, 38,89% (28) appear in any 4-gram, showing a slightly lower rate in comparison to the number

of 2-grams appearing in a 3-gram. One of the 72 3-grams stands out: "human – computer – interaction", given that it appears in a total of eight 4-grams, which are "based – human – computer – interaction", "research – human – computer – interaction", "human – computer – interaction – system", "human – computer – interaction – analysis", "human – computer – interaction – hci", "learning – human – computer – interaction", "human – computer – interaction – interface" and "method – human – computer – interaction".

Only 35.35% (35) of the 99 presented 4-grams appear in any of the 5-grams, with a maximum of two appearances. From these, it is worth mentioning the first 4-gram, "functional – near – infrared – spectroscopy", which, in addition to having the highest number of occurrences (13), also appears in two 5-grams. For the 5-grams, 42,86% (9) of them appear in some of the 6-grams. From this point forward, it is possible to notice that the subjects become progressively more specific due to the higher number of words in the *n*-grams. Only two of the nine 5-grams appear in a maximum of two different 6-grams, while the others appear in only one.

The four first 6-grams appear to be related to a single common topic, which, in turn, was covered in four different publications of the HCII. This becomes more evident when considering that four from all six 6-grams (66,67%) appear in one of the 7-grams. All the three different 7-grams refer to the same very particular topic, given that all three share the same words.

4.2 IHC

A total of 713 articles were identified for this symposium, which were published in 18 editions between 1998 and 2019 (*M* = 39,61; *SD* = 17,75). In total, 8.512 words compose the set of all words from the titles of these publications from IHC.

4.2.1 IHC word analysis

The most significative words of the IHC were identified based on the set of all words from its titles (8.512) minus the stopwords, being **N4 the set of all different significative words from IHC**, i.e., 2.088 words.

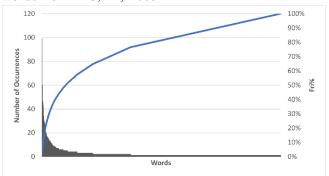**Figure 7.** Histogram of the N4 (2.088) words from IHC.

Figure 7 shows a histogram of the distribution of N4 by its absolute frequency in descending order. The percentages of cumulative relative frequency (*Fri%*) are 12,98% (n = 15), 25,04% (n = 44), 50,05% (n = 208) and 75,00% (n = 732), which indicate that the most recurring significative words are also concentrated in the first positions of the histogram,

being **n4 the set of the first *n* words with highest *fi* from N4 and with *Fri%* equal to approximately one fourth of the total**.

The word correlation analysis of the IHC was performed by creating the word correlation matrix of this symposium, with a total of 44 dimensions (Figure 8), since the 44 (n4) most recurring words from the 2.088 (N4) significative words from IHC concentrate, alone, approximately 25% of the total of occurrences. Thus, this matrix presents the 990 possible correlations between the 44 most recurring words from IHC, which have 17 or more occurrences. As described

in in Section 3.4.1, words from different languages may appear in this correlation matrix.

The ranking of words from the IHC is given by any axis of this matrix. The most recurring words that also have more than 40 occurrences are (in descending order): "design" (#1; 99), "web" (#2; 60), "interfaces" (#3; 60), "interaction" (#4; 52), "study" (#5; 48), "interação" (#6; 46), "user" (#7; 46), "avaliação" (#8; 44), "interface" (#9; 44) and "evaluation" (#10; 43). The analysis of this ranking reveals, in a sense, the most covered topics in this symposium.
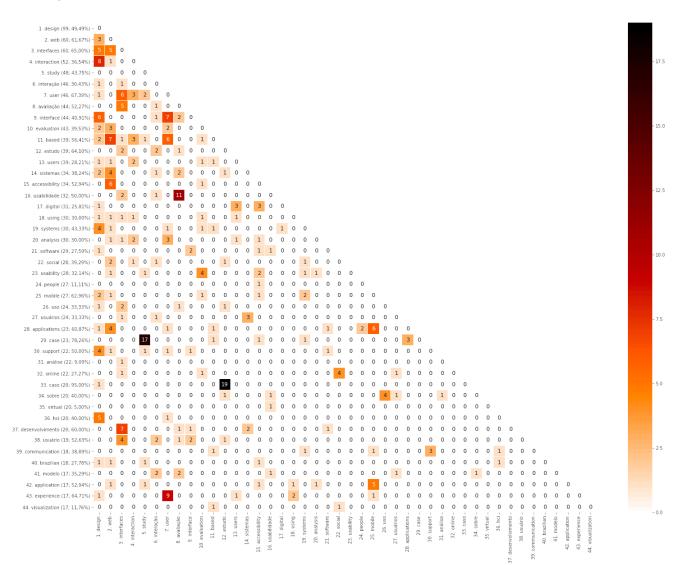

Figure 8. Correlation matrix of the n4 (44) most recurring words from IHC.

As with the previous analysis, the correlation of a word with one another in this matrix can contribute to reveal new subjects that are covered in the conference. The correlations between two words in this matrix that have the highest number of occurrences (and number of occurrences ≥ 3), are, in descending order: "estudo ↔ caso" (#12,33; #33,12; 19), "study ↔ case" (#5,29; #29,5; 17), "avaliação ↔ usabilidade" (#8,16; #16,8; 11), "user ↔ experience" (#7,43; #43,7; 9), "design ↔ interaction" (#1,4; #4,1; 8), "web ↔ based" (#2,11; #11,2; 7), "interfaces ↔ desenvolvimento" (#3,37; #37,3; 7), "interface ↔ user" (#9,7; #7,9; 7), "accessibility ↔ web" (#15,2; 6), "mobile ↔ applications" (#25,28; #28,25; 6), "hci ↔ design" (#36,1; 5), "application ↔ mobile" (#42,25; 5), "evaluation ↔ usability" (#10,23, #23,10; 4), "sistemas ↔ web" (#14,2; 4), "systems ↔ design" (#19,1; 4), "social ↔ online" (#22,32; #32,22; 4), "uso ↔ sobre" (#26,34; #34,26; 4), "support ↔ design" (#30,1; 4), "usuário ↔ interfaces" (#38,3; 4), "users ↔ digital" (#13,17; #17,13; 3), "digital ↔ accessibility" (#17,15; 3), "analysis ↔ user" (#20,7; 3), "usuários ↔ sistemas" (#27,14; 3) and "communication ↔ support" (#39,30; 3). These other correlations of words reveal that

many works presented at the symposium describe results of case studies. There is also a recurrence of works that explore issues related to the usability of interactive systems, and several works based on web platforms, especially regarding concerns about web accessibility. The mobile platform also appears prominently.

4.2.2 Analysis of terms from IHC

A total of 429 titles in English were identified from the IHC, based on a total of 713 titles from publications and the method described in Section 3 (especially the description in Section 3.4.2). The other titles were either in Portuguese (280) or in Spanish (4).

The terms of this symposium were identified from the total of 429 titles of its publications in English, being **N5 the set of all different terms in English from IHC**, i.e., 1.197 terms. Figure 9 shows the distribution of N5 in a histogram of its absolute frequency in descending order. The percentages of cumulative relative frequency (*Fri%*) are 12,80% (n = 21), 25,11% (n = 79), 50,03% (n = 377) and 75,02% (n = 787). The most recurring English terms from IHC are also concentrated in the first positions of this histogram, being **n5 the set of the first *n* terms with highest *fi* from N5 and with *Fri%* equal to approximately one eighth of the total**. It is worth mentioning that the 22nd, 23rd, 24th and 25th terms with highest absolute frequency had

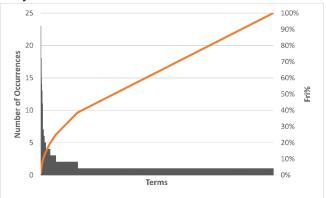exactly the same absolute frequency of the 21st term. Therefore, the first 25 terms were considered for this analysis.

Figure 9. Histogram of the N5 (1.197) terms from IHC.

The analysis of correlation between the terms of the IHC was performed through the creation of the term correlation graph of this symposium. This graph has a total of 25 vertices (Figure 10), since the 25 (n5) most recurring terms within the total of 1.197 (N5) terms from IHC concentrate, alone, over 12,5% of all occurrences. Thus, this graph presents the 103 connections (edges) between these 25 vertices that represent the most recurring terms from IHC, each having five or more occurrences.
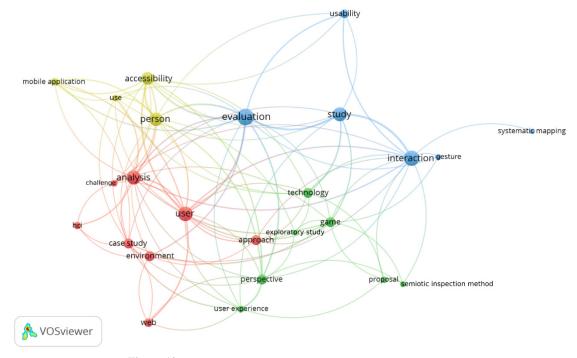
Figure 10. Correlation graph of the n5 (25) most recurring terms from IHC.

In this term correlation graph, the terms were grouped into four clusters (the different colors of the vertices), which are described next by descending order of occurrences. The cluster C1, in red, is composed of 8 terms: "user" (32), "analysis" (29), "approach" (16), "case study" (16), "environment" (16), "web" (12), "hci" (8) and "challenge" (7). C2, in green, consists of 7 terms: "game" (16), "perspective" (15), "technology" (15), "proposal" (7), "user experience" (7), "semiotic inspection method" (6) and

"exploratory study" (5). C3, in blue, consists of 6 terms: "evaluation" (42), "interaction" (35), "study" (29), "usability" (12), "gesture" (6) and "systematic mapping" (5). C4, in yellow, is composed of 4 terms: "person" (27), "accessibility" (24), "mobile application" (8) and "use" (7).

In this graph, some vertices were formed by terms composed of two or more words, thus showing the strong relation between these words and revealing new recurring topics (presented in descending order of occurrence): "case

study" (16), "mobile application" (8), "user experience" (7), "semiotic inspection method" (6), "systematic mapping" (5), "exploratory study" (5). The majority of terms refers to topics that were already identified through the word correlation matrix. However, a total of 13 new terms that appeared in this analysis were not present as words in the word correlation matrix, such as "systematic mapping" and "semiotic inspection method", and therefore indicate new recurring topics. These terms are the following, presented in descending order of occurrence: "person" (27), "approach" (16), "environment" (16), "game" (16), "perspective" (15), "technology" (15), "challenge" (7), "proposal" (7), "use" (7), "gesture" (6), "semiotic inspection method" (6), "exploratory" de "exploratory study" (5) e "systematic mapping" (5). The appearance of these terms shows other topics that are explored, for example, in studies involving gestural interactions and games.

4.2.3 Analysis of sequences of words from IHC

As with the HCII, the *n*-grams of the IHC were generated based on the 713 titles from its publications, being **N6 the set of all different instances of a given *n*-gram from IHC**. The N6 of the different *n*-grams from IHC are: 2-gram (307), 3-gram (69) and 4-gram (24). Each N6 was represented in a histogram based on its absolute frequency in descending order (Figure 11). Table 2 shows the different cumulative relative frequencies in percentage (*Fri%*). The histograms of the different *n*-grams from IHC are also less concentrated in its first positions, and they tend to become more homogeneous as the number of words in sequence (*n*) increases. For this analysis, **n6 is the set of the first *n* instances of a given *n*-gram with highest *fi* from N6 and with *Fri%* equal to approximately one eighth of the total**.

It is worth mentioning that the sample size (N6) of each different *n*-gram from IHC varies, and therefore n6 also varies according to the rules defined in the method (details in Section 3.4.3). It is also worth mentioning that the value of n6 was defined so that its *Fri%* is approximately one eighth of the total, and that $n \geq 3$.

The analysis of sequences of words of the IHC was performed through the creation of the *n*-grams correlation network for this symposium (Figure 12). Based on the data from IHC, and considering the method described in Section 3, the network connects three different sets of *n*-grams, i.e., 2-gram, 3-gram and 4-gram. Also, as described in Section 3.4.1, words from different languages may appear in Figure 12.

Table 2. N6 and n6 of the different *n*-grams from IHC.

n-gram	*Fri%*				N6	n6
	12,5%	25%	50%	75%		
2-gram	12 (13,09%)	30 (25,08%)	93 (50,28%)	194 (75,14%)	307	17
3-gram	4 (14,29%)	10 (26,19%)	27 (50,00%)	48 (75,00%)	69	4
4-gram	3 (16,00%)	6 (28,00%)	12 (52,00%)	18 (76,00%)	24	2

Considering the 307 (N6) possible 2-grams from IHC, the instance in the 12th position of the set has *fi* = 7. Since the bigram must present all instances with this same *fi*, it was generated up to the 17th position (Figure 12a). For the 69 (N6) possible 3-grams from IHC, the instance in the 4th position of the set has *fi* = 5. As the trigram must present all instances with this same *fi*, it was generated up to the 4th position (Figure 12b). Finally, considering the 24 (N6) possible 4-grams from IHC, the instance in the 3rd position of the set has *fi* = 2. Since the quadgram should present all instances with *fi* $\geq$ 3, it was generated up to the 2nd position (Figure 12c), as *fi* $\leq$ 2 starting from the 3rd 4-gram.

For the IHC, there is only one case of n-gram that is not derived from the (n-1)-gram and does not appear on the (n+1)-gram, which is the case of the 3-gram "human – computer – interaction" (#4; 5), i.e., the English version of the most recurring 3-gram of the IHC, "interação – humano – computador" (#1; 7). Still in the IHC, there is also a case of n-gram that does not appear in the (n+1)-gram, but appears in a following n-gram: the 2-gram "desenvolvimento – interfaces" (#7; 7), which does not appear in any 3-gram, but appears in the 4-gram "desenvolvimento – interfaces – homem – computador" (#1; 3).
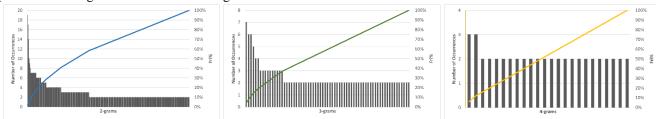

Figure 11. Histograms of the *n*-grams from IHC: (a) bigram; (b) trigram; (c) quadgram.

In the *n*-grams correlation network from IHC, it is possible to notice that two 2-grams converge to the same 3-gram: "humano – computador" (#6; 8) and "interação – humano" (#10; 7), which appear simultaneously in the 3-gram "interação – humano – computador" (#1; 7). By the number of occurrences of these *n*-grams, it can be noticed that these two 2-grams are related exclusively to this 3-gram, with the exception of "humano – computador", which has

one more occurrence related to other instance that does not appear in the set of 3-grams.

A recurring phenomenon that appears in the IHC network are *n*-grams that do not appear in any (*n*+1)-gram or forward, thus indicating that more general and relevant topics dissolve into more specialized ones with less occurrences. In the set of 2-grams, for example, this phenomenon occurs in the majority of the elements: "estudo

– caso" (#1; 19), "case – study" (#2; 17), "avaliação – usabilidade" (#4; 10), "user – experience" (#5; 9), "sistemas – colaborativos" (#8; 7), "social – networks" (#9; 7), "systematic – mapping" (#12; 7), "interaction – design" (#13; 7), "método – avaliação" (#14; 7), "web – based" (#15; 7), "user – interface" (#16; 7) and "tv – digital" (#18; 8).

From these 2-grams, only 23,53% (4) appear in a 3-gram, which are "visually – impaired", "humano – computador",

"interação – humano" and "semiotic – inspection". It is interesting to notice that the 2-grams "interação – humano" and "humano – computador" are associated with the same 3-gram, "interação – humano – computador", the main topic of the conference. None of the four 3-grams of the IHC appears in any of the 4-grams. Finally, the following 4-grams were identified: "desenvolvimento – interfaces – homem – computador" (#1; 3) e "brazilian – symposium – human – factors" (#2; 3).
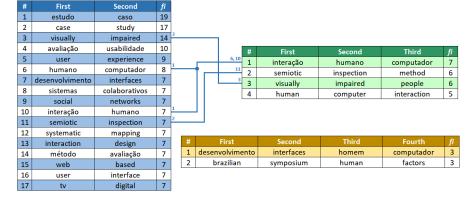
#	First	Second	fi
1	estudo	caso	19
2	case	study	17
3	visually	impaired	14
4	avaliação	usabilidade	10
5	user	experience	9
6	humano	computador	8
7	desenvolvimento	interfaces	7
8	sistemas	colaborativos	7
9	social	networks	7
10	interação	humano	7
11	semiotic	inspection	7
12	systematic	mapping	7
13	interaction	design	7
14	método	avaliação	7
15	web	based	7
16	user	interface	7
17	tv	digital	7

#	First	Second	Third	fi
1	interação	humano	computador	7
2	semiotic	inspection	method	6
3	visually	impaired	people	6
4	human	computer	interaction	5

#	First	Second	Third	Fourth	fi
1	desenvolvimento	interfaces	homem	computador	3
2	brazilian	symposium	human	factors	3

Figure 12. *n*-grams correlation network from IHC: (a) bigram; (b) trigram; (c) quadgram.

4.3 IHC x HCII

First, it is important to highlight that since the statistical tests allow the comparison only between words that appear in both conferences, as described in Section 3, there are 25 cases of words that, although having many instances in HCII, do not appear in IHC. Words such as "display" and "robot", which appear between the first 100 significant words of HCII, in addition to other words such as "chinese", "image", "driving", "workload" and "vehicle", which appear considering 50% of the *Fri%* of HCII (first 276 words), were not comparatively analyzed as they have no instance in IHC.

The inter-conference analyses will be presented in an ascending order of instances, i.e., going from the first most significant words of each conference to a more general analysis, considering all significant words of both conferences. The tests were performed considering a significant value of 5% (p < 0,05) and a 95% confidence interval. The scale defined in Koo and Li (2016) was employed for the agreement values, in which values lower than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and higher than 0.9, are considered poor, moderate, good, and excellent agreements, respectively.

4.3.1 Correlation and agreement between the rankings of IHC and HCII

In the more specific analysis of the main topics of the conferences, considering only the significant words that appear in the 100 first positions of each ranking (with a total of 134 words, considering the intersections), an ICC value of -0,076 (p = 0,665) was obtained, which shows that there is no significant agreement between the two conferences. This means that, considering the current sample, there is no significant trend for the same words to be positioned in close or equal positions, i.e., each conference tends to assign a

different order of importance to each topic, which may be higher or lower. In relation to the correlation, a correlation coefficient with value τb = 0,300 (p < 0,01) was obtained, showing a slight positive correlation. The scatterplot in Figure 13 shows this distribution, highlighting the most significative words that are positioned further apart between the rankings. The lines in red represent the 100th position of the ranking of each conference.
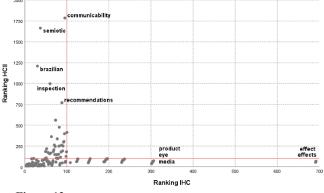

Figure 13. Scatterplot of the first 100 words of the ranking of each conference.

The figure shows that the conferences are quite different in relation to their most covered topics, since words that appear in higher ranking positions in the IHC (first 100 positions), such as "semiotic", "communicability", "inspection" and "brazilian", appear in much lower positions in the HCII (over the 1000th position), showing that the IHC community considers these topics as more prominent in comparison to the HCII. On the other hand, words such as "effect(s)", "product" and "media" appear in much lower positions in the HCII in comparison to the IHC, also showing some of the main differences between the conferences. The correlation shows that there is a positive association between the two rankings, i.e., although there is a degree of discordance in regard the importance of each topic given

their positions in the ranking, there is also a slight tendency of the higher the position of a word in one of the rankings, the higher its position in the other one, and vice-versa.

As Figure 13 shows, this slightly positive correlation is given by the first words of both rankings (lower left portion of the Figure), which are more concentrated, and is lower in value due to the dispersion of the outliers along the axes. The scatterplot in Figure 14 shows a cutout of this graph (Figure 13), with the concentration of words in common between the two conferences up to the 100th position (a total of 63 words), and the regression line highlighted in red. In this Figure, it is possible to notice that although there is a positive linear trend in the ranking of words of both conferences, there is also an increasing spacing between the words (which is even larger considering the entire sample), partially showing the results of both the correlation and the concordance tests.
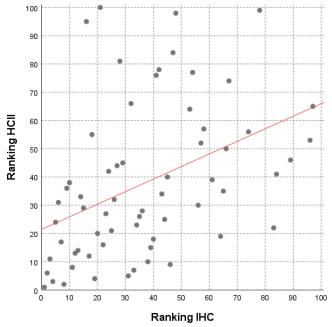

Figure 14. Scatterplot of the words in common between the two conferences up to the hundredth position of their rankings.

In a broader analysis, considering 50% of the *Fri*% of each conference, i.e., 276 words from HCII and 208 words from IHC (with a total of 312 words, including the interceptions), an ICC value of -0.046 was obtained (p = 0.653), showing that there is also no significant agreement between the two conferences considering a larger sample size. For the correlation, a coefficient value of $\tau b = 0.255$ was obtained (p < 0.01), which is also a slightly positive, statistically significant correlation, but weaker than the previous analysis. The scatterplot in Figure 15 shows the distribution of words in both rankings. The Figure also highlights words positioned farther between the two conferences. The red horizontal line represents the 276th position of the HCII's ranking, while the vertical one represents the 208th position of the IHC's ranking.

Figure 15 shows, in addition to words from the previous analysis, new outliers, such as "diagrams", "contributions" and "molic", which are positioned with a much higher importance in the IHC in comparison with its counterpart in HCII. In the opposite direction, words positioned in a more

distant position in the IHC ranking, such as "vision", "concept", "sensor" and "automated", appear with more importance in the HCII, showing a more striking difference between the two conferences.

Considering the entire sample, consisting of all words that appear in both conferences (1.397 words in total), the analysis of agreement revealed an ICC value of 0.224 (p < 0.01), with a confidence interval of 95% between 0.070 and 0.347, which represents a poor agreement, i.e., although there is a certain degree of agreement, similar words tend to be in different and distant positions in relation to each ranking. The correlation test, on the other hand, revealed a coefficient value of $\tau b = 0,711$ (p < 0,01), which indicates a strong positive correlation, i.e., words placed in lower positions in the ranking of one conference also tend to be in a lower position on the other.
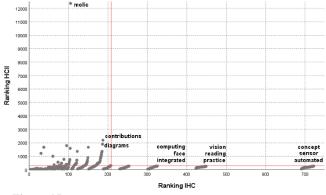

Figure 15. Scatterplot considering 50% of the *Fri*% of the ranking of each conference.

Regarding this comparative analysis of the conferences, the results of the ICC show that, although the two conferences have a small degree of agreement in general, they differentiate in relation to their main and specific topics. A weak concordance was found only when considering the set of all words in common between them. The positive correlation, in this case, can be explained by the large number of words with only one instance or with very few instances in both conferences, which tend to be positioned in similar positions in the lower end of both conferences' rankings.

The exceptions are the outliers, which have a very high position in one of the ranking, but a lower position in the other, and show the main differences between the importance that each conference attributes to these specific topics. These differences became even more evident with the specific analysis, considering only the first 100 words and 50% of the *Fri*% of each conference, since there is no significant agreement, and the correlation is considerably weaker in both cases. The large number of outliers explains both results, since each extreme value tends to lower both metrics, and corroborates with the significant difference between the most covered topics in the two conferences. Another factor that has influence on this difference is the presence of words that do not appear in the IHC, but have a high classification in the HCII, such as "robot" and "chinese", as aforementioned.

5   Discussion

The choice of the HCII as one of the conferences to be investigated in this study was motivated, as previously described by the results presented in Gasparini et al. (2017). The authors of the present study do not have the pretension of discussing the method of selection of the articles that are published and presented at the HCII. The authors also have no intention of understanding, considering only the results of this study, the phenomenon associated with the fact that HCII is the conference where Brazilian researchers of the field of HCI most publish their work after the IHC itself. We understand that the results of this study may help in the discussions that have already started in related work, and may also contribute to the Brazilian scientific community in the field of HCI to reflect on their paths.

The main findings that were obtained from the data were presented in Section 4. It is worth mentioning that it is possible for the reader to use the data and the visualizations presented in this work to investigate other questions that were not explored in the present study, such as the progress over time of specific topics and their individual importance for each conference. The method defined for this study had as its basic principle to be as impartial as possible, aiming to mitigate possible biases regarding the description of the presented topics. As such, all results were described after a rigid processing of the data with criteria that were defined previously. The most frequent words and terms were highlighted in an objective manner in all generated artifacts. For example, the analysis of the n-grams correlation network and the presented examples are intended, exclusively, to make explicit the sequence of words between *n*-grams and to motivate the reader to search for new sequences of words in the n-grams network. The reading of the *n*-grams network in its various "paths" shows the importance of this combined analysis that the n-grams offer, which goes far beyond just correlating words.

Thus, the frequency analysis contributed to the better visualization of the most frequently explored topics in the publications, while not discarding the importance of topics that haven't been much explored in the literature (yet). However, to analyze a conference with tens of thousands of publications lacks an objective strategy to better visualize its propensities, and the frequency-based analysis helps in this regard. Its exceptions may also be identified by searching for the "outlier" topics.

Considering that some visualizations were produced manually by the authors of this work, which increases the risk of human error in the process, the development of tools to automatically generate the word correlation matrix and the *n*-grams correlation network were opportune. Thus, a possible future work would be the development of a tool focused exclusively on the automated generation of these two visualizations, that is, a tool capable of automatically processing a given input data (e.g., a file structured in plain text) and generating its visualizations according to the method described in Section 3, and therefore being less prone to human error.

In this sense, we understand that the present study contributes not only with the analysis and comparison between the two conferences of the field of HCI, but also with the bibliometric methods that were employed to perform both the individual and the comparative analysis of the conferences. Certainly, with a few adaptations, these methods can be employed not only in bibliometric analyses, but also in big data in general.

5.1 Limitations and threats to validity

As threats to the validity of this study, it is worth mentioning that, due to limitations of the VOSViewer in relation to the supported language for the textual analysis, the set of titles from the IHC had to be filtered. Only titles that were in English were processed by the software; therefore, titles in Portuguese or Spanish were excluded from this analysis.

In addition, as described in Section 4.3, the statistical analysis was performed considering only words that appeared in the titles of works within both conferences; therefore, words that appeared in the titles of works from only one of the two conferences were excluded from the statistical analysis. Still in relation to this analysis, it is also important to highlight that the process of comparison involved a step of translating the words from the titles in Portuguese and Spanish to English. Although each word was translated manually and individually to avoid semantic errors produced by automatic translation software, it is possible that not every translated instance corresponds accurately to its actual translation, due to both human error and the original context from which the word was isolated (i.e., specific words may have different translations depending on the semantic context of the title, which is not always clear considering only one isolated word). However, the titles were kept in their original languages for the word correlation matrix and the *n*-gram correlation network, given that the sequences of words in the titles could be altered in the translation process.

6   Conclusion

Characterizing the main interests of scientific communities has been an important way to understand what they value and where they are heading in terms of research. In this work, we analyzed two conferences from the field of Human-Computer Interaction: the HCII, which is the international conference where the Brazilian researchers most publish their works (c.f. Gasparini et al. 2017), and the IHC, which is the most important conference of the field in the Brazilian context. For this study, more than 209 thousand words from over 18 thousand publications of both conferences were analyzed using various statistical and visualization approaches.

The results indicate that the Brazilian scientific community of the field of HCI have been exploring topics that are also of interest for the international community, although it also has very different priorities in relation to the covered topics, as presented and discussed in more detail in this study. In addition, both analyzed conferences presented exclusive research topics of interest, which become evident

both numerically through the statistical analysis and by investigating the outliers, which represent topics that are given more importance in one conference in relation to the other.

As future works, we suggest an analysis based on clusters of words to better explore other dimensions of their relations, and to compare the IHC with other important international conference of the field. Furthermore, the method developed and employed in this work for its many analyzes can be adapted for other contexts, including future bibliometric studies and the processing of large volumes of data. Further developing those methods can provide new ways of processing and visualizing the collected data.

Acknowledgements

References

Aghimien, D.O., Aigbavboa, C.O., Oke, A.E., & Thwala, W.D. (2019). Mapping out Research Focus for Robotics and Automation Research in Construction-related Studies: A Bibliometric Approach. *Journal of Engineering, Design and Technology*, 18(5):1063–1079.

Baranauskas, M.C.C., & Souza, C.S. (2006). Desafio nº 4: Acesso Participativo e Universal do Cidadão Brasileiro ao Conhecimento". *Computação Brasil*, 7(27):7.

Baranauskas, M.C.C., Souza, C.S., & Pereira, R. (2015). I GranDIHC-BR Grand Research Challenges in Human-Computer Interaction in Brazil.

Barbosa, S.D.J., Silveira, M.S., & Gasparini, I. (2017). What Publications Metadata Tell Us about the Evolution of a Scientific Community: The Case of the Brazilian Human-Computer Interaction Conference Series. *Scientometrics*, 110(1):275–300.

Buchdid, S.B., & Baranauskas, M.C.C. (2012). IHC em Contexto: O que as Palavras Relevam sobre Ela. 11th Brazilian Symposium on Human Factors in Computing Systems (IHC'12), ACM, 199–208.

Buchdid, S.B., & Baranauskas, M.C.C. (2014). Is There HCI in iDTV? An Exploratory Study on Their Words. 16th International Conference on Human-Computer Interaction (HCII'14), Springer, 47–57.

Bueno, A.O., Ferreira, L.C., Ferreira, V., & Anacleto, J.C. (2016). Research Trends in HCI in Brazil: An Analysis in Relation to the GranDIHC-BR. 15th Brazilian Symposium on Human Factors in Computing Systems (IHC'16), ACM, 10 pages.

Coelho, T., Barbosa, G.A.R., Silva, I.S., Coutinho, F.R.S., & Silva, F.R. (2017). An Overview of Researches on Digital Accessibility before and after the Great Challenges of SBC 2006-2016: A Systematic Literature Review. XVI Brazilian Symposium on Human Factors in Computing Systems (IHC'17), ACM, 10 pages.

Gasparini, I., Mendonça, F.C., Silveira, M.S., Barbosa, S.D.J., & Schroeder, R. (2017). Crossing the Borders of IHC: Where Else Have Our Researchers been Publishing? 16th Brazilian Symposium on Human Factors in Computing Systems (IHC'17), ACM, 10 pages.

Granatto, C.F., Pallaro, M.A.P., & Bim, S.A. (2016). Digital Accessibility: Systematic Review of Papers from the Brazilian Symposium on Human Factors in Computer Systems. 15th Brazilian Symposium on Human Factors in Computing Systems (IHC'16), ACM, 10 pages.

Gupta, A. (2015). Five Years of IndiaHCI: A Scientometric Analysis. 7th International Conference on HCI (IndiaHCI'15), ACM, 56–61.

Koo, T.K., & Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global Maps of Science based on the New Web-of-Science Categories. *Scientometrics*, 94:589–593.

Lima, F.M.C., Miranda, L.C., & Baranauskas, M.C.C. (2018). Two HCI Communities in Perspective: Revealing and Contrasting Trends. 17th Brazilian Symposium on Human Factors in Computing Systems (IHC'18), ACM, 10 pages.

Lima, F.M.C., Miranda, L.C., Mendes, G.A.V., & Baranauskas, M.C.C. (2019). An Intra and Inter-conferences Correlation Analysis of the Field of HCI: Revealing New Trends from the International and Brazilian Communities. 18th Brazilian Symposium on Human Factors in Computing Systems (IHC'19), ACM, 11 pages.

Liu, Y., Gonçalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994-2013: Mapping Two Decades of Intellectual Progress through Co-word Analysis. 32nd SIGCHI Conference on Human Factors in Computing Systems (CHI'14), ACM, 3553–3562.

Mubin, O., Al Mahmud, A., & Ahmad, M. (2017). HCI Down Under: Reflecting on a Decade of the OzCHI Conference. *Scientometrics*, 112:367–382.

Padilla, S., Methven, T.S., Corne, D.W., & Chantler M.J. (2014). Hot Topics in CHI: Trend Maps for Visualising Research. *Extended Abstracts on Human Factors in Computing Systems*, ACM, 815–824.

Santana, V.F., Neris, V.P.A., Rodrigues, K.R.H., Oliveira, R., & Galindo, N. (2017). Activity of Brazilian HCI Community from 2012 to 2017 in the Context of the Challenge 'Future, Smart Cities, and Sustainability'. XVI Brazilian Symposium on Human Factors in Computing Systems (IHC'17), ACM, 5 pages.

Silva, E., Torres, B., Sacramento, C., Capra, E.P., Ferreira, S.B.L., & Garcia, A.C.B. (2018). Privacy: What is the Research Scenario in Brazilian Symposium IHC? 17th Brazilian Symposium on Human Factors in Computing Systems (IHC'18), ACM, 8 pages.

Steinmacher, I., Silva, M.A.G., Côgo, F.R., Chaves, A.P., & Gerosa, M.A. (2013). Tópicos de Pesquisa e Rede de Coautoria no Simpósio Brasileiro de Sistemas Colaborativos. X Brazilian Symposium in Collaborative Systems (SBSC'13), SBC, 88–95.

Van Eck, N.J., & Waltman, L. (2010). Software Survey: VOSviewer, A Computer Program for Bibliometric Mapping. *Scientometrics*, 84:523–538.

Van Eck, N.J., & Waltman, L. (2013). VOSviewer manual. *Leiden: Univeristeit Leiden*, 1(1):1–53.