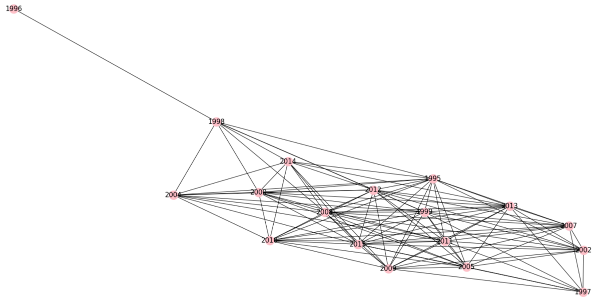
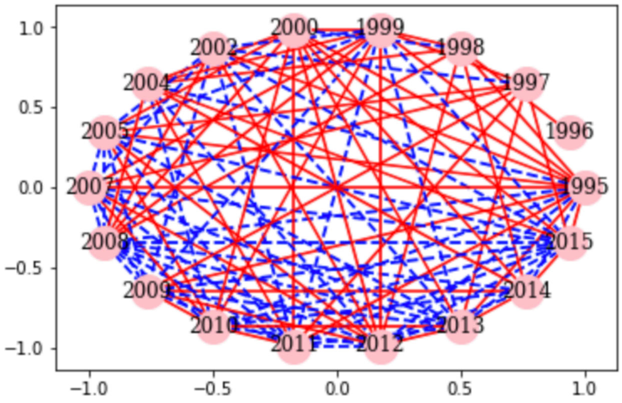
Learning Science Trends Across Two Decades

Sophie Mann, Harvard Graduate School of Education, smmann@gse.harvard.edu





Abstract: This analysis serves to elucidate some of the trends that have waxed and waned across the past two decades of International Society of Learning Sciences conference proceedings. Present findings indicate that an increased focus on social emotional learning and technology-centric trends align with a rapidly growing focus on technology in all sectors, especially in the field of education, as well as the major influx of focus in social emotional learning. This was demonstrated in great increases in means of such words across decades, as well stronger connections in a network analysis of the later decade years between each other, in comparison to how similar their vocabularies were to older years, which had fewer instances of technology-based or SEL-driven words.

Introduction

The International Society of the Learning Sciences (ISLS) presents findings in the field biannually at the Conference on Computer-Supported Collaborative Learning (CSCL) and at the ISLS proceedings, as well. Throughout the past two decades, emerging trends from across education and the learning sciences demonstrate how these fields change, as well as what topics remain at the forefront of researchers' and practitioners' exploration. These proceedings receive insights and findings from learning science frontrunners from across the globe, creating a nuanced look at the field. Presently, the past 20 years' abstracts and proceedings in their entirety exist for people to inform their praxis.

The analysis of this corpus, therefore, can be used by other practitioners, researchers, and learning scientists to inform classroom teaching, out-of-school time learning experiences, and future research in the ever-expanding area of the learning sciences. Further, because these abstracts come from both the learning sciences and, more specifically, the computer-supported learning community the topics discussed are interdisciplinary in nature with a distinct focus on collaboration. In turn, this works to support educators of all types in increasing the social-emotional and collaborative natures of their curricula, experiments, and interactions with students and peers moving forward. Social emotional learning in particular has been of increased prevalence across the last decade in the field of education, as it has been found to increase students' academic achievement (Panayiotou, 2019; McCormick, 2015). Additionally, adaptive learning and technology-assisted pedagogy has become more prevalent as practitioners begin to become more comfortable with technology in the classroom, as compared to 1995-2005, especially as it proves to help student outcomes, as well. (Ross, 2018; Szijarto, 2018); With the most recent proceedings in the corpus in 2015, this analysis and the trends across the past two decades can make further proceedings' findings situated in the most persistent trends of recent years. In the same vein, the analysis conducted on large corpuses is applicable to other subjects as well (Goncalves, 2014; Chen, 2016), including human-computer interactions, economics, computer engineering, and others.

Dataset

The dataset presented here includes the abstracts from the last two decades of ISLS and CSCL conference papers. This included 1995-2015, excluding 2003, 2006, and 2011. The abstracts were in text files by year and each individual abstract within a year's text file was denoted by line spaces. In order to glean trends and understanding from year to year, a cosine similarity matrix was created to compare the Top 25 words throughout the entire corpus' frequency delineated by year and each year's similarity to one another. This was incredibly important, as this allowed for further analysis of how strongly each year related to other years through degree centrality and betweenness centralities, giving insight into the changing trends across the years.

Degree centrality demonstrates how connected certain nodes are to other nodes within a network. Here, this allowed us to determine which years were most connected and most strongly to one another. Similarly, betweenness centrality is how much control the node has over the network in that it has the largest number of shortest paths between it and other nodes. In the case of this dataset, degree centrality indicates which nodes are most similar to one another. Betweenness centrality indicates first-degree connections between nodes.

Research Questions

The questions posed here were two-fold. First, what trends persisted and which waned across the two decades? Did researchers begin moving forward more technologically inclined trends as technology steadily advanced throughout the 2000s? Secondly, did social emotional learning become more preeminently discussed in the later decade? The first question touches upon a general world trend of technological advancement, while the second looks at the increased desire to teach students "21st Century skills" or "soft skills," like self-awareness or responsible decision making.

Hypotheses

Because the conferences focus around collaborative learning and the learning sciences, I anticipate that many of the most frequently discussed topics will continue to be of importance across the decades with heavier focuses on technology, as technological advancement increases. learning, artificial intelligence and its pertinence to education, and the importance of personalized learning, which often involves technology, have been more prominent across the field in the past decade, which is why technology-oriented words may become more frequently used. Similarly, social-emotional learning has also become more prominent, which is words like "social" may also increase across the decades. It can be assumed that years in the later 2000s will be most similar to one another, while words in the earlier years will be most similar to one another, based on these trends. For example, it is anticipated that 1995-1999 will be more similar to one another than 1995 and 2015, and 2010-2015 will be more similar to each other, as well.

Methods

In order to clean the data, all abstracts were compiled into a text file. Subsequently, the abstracts were turned into a dictionary of conference years and their respective abstracts. From this, all punctuation and digits were removed to ensure an easier analysis process. A word frequency function was constructed to remove all words that were not meaningful, such as articles, conjunctions, and abbreviations, such as CID which was a language tag. These were labeled as stop words and removed from all of the abstracts. The most frequent words across the entire corpus were then recalculated. The Top 25 words from across the corpus were then checked across each year, and these frequencies were compiled into a Pandas DataFrame.

From this matrix, the cosine similarities were calculated to see how similar each year's vocabulary frequencies were. Through this, each year's frequencies were vectorized, normalized, and the cosine between their vectors was calculated. Cosine similarity demonstrates how similar vectors are even if they are not close to one another, though most of these vectors have a very high cosine similarities. This allowed for network analysis, further data visualization using plotly_express, and Seaborn.

Results

Some data cleaning revisions were required prior to moving forward. For example, CID and CSCL were in the top 25 words. However, CID does not stand for anything of note, when the actual texts of the abstracts and papers were examined, and CSCL is the name of the conference. These were subsequently removed, along with stray numerals, and punctuation was cleaned a second time. After looking through the list, it made sense to remove words that were not indicative of trends, such as "education," which appeared many times because of the nature of the papers. I also removed words like "findings" that are very common in papers in general. A matrix was created with the Top 25 words with totals. These were summarized in a word cloud (Figure 1).

After completing further cleaning, the cosine similarities of the frequencies data frame were calculated. This indicated the similarity between years, demonstrating how similar the trends in the vocabulary remained. To demonstrate this visually, a heatmap was created using Seaborn (Figure 2). The heatmap shows a fairly high similarities between each year, which makes sense as these conferences focused heavily on collaborative learning. However, some of the years were more similar than others, which was found by sorting the Top 25 words df for highest similarities. The Top 20 most similar years can be found in Table 1.

This was then represented using network analysis. Degree and Between Centrality were calculated and represented using the NetworkX library. They can be found in Figures 3 and 4 respectively. These demonstrate how connected each year is to other years. The least connected appears to be 1996, as well as having the small number of shortest paths, whereas the 1999-mid-2000s have the most connections and short paths. This makes sense, considering the strongest connections according to Table 1 are in the years represented with the highest degree centralities, meaning they have the most similar vocabularies. Additionally, 1996 being least densely connected makes sense when looking at the Top 25 words in 1996 vs. other years; it appears as though there are fewer abstracts total, as the word frequencies across the Top 25 are the lowest of any year.

In order to represent this, the cosine similarities under .80 were removed. This threshold was set to demonstrate the strongest connections within the network. Without removing the least strong connections, all years were connected to one another, giving a density score of 1.00 and providing no meaningful results, especially for the earlier years which typically have the lower connectivities, as is seen here.

Following this analysis and visualization, the Plotly Express library, a visualization library that allows for interactive graphical representations of data, was used to demonstrate how similar years were visually. In Figure 5, each year's similarities are demonstrated, as they are in the cosine similarity matrix, with varying colors showing different similarity strengths. Hovering over there each year, as indicated, shows the years represented and their similarity. Unsurprisingly, two of the earliest years (1996 and 1997) are least similar to 2007 and 2014. What did come as a surprise was that these two years are also very dissimilar to one another. Other interesting findings can be elucidated from this graph in particular which will be pursued in further sections.

Topic modeling methods were used to elucidate which words may go together most readily. Three clusters were chosen for kmeans based on an inertia plot. Words that most prominently went together are indicated here:

Cluster 0	Cluster 1	Cluster 2
practices interaction information inquiry data teacher online development tools process group activities social activity community analysis technology learners project collaboration environment computer support collaborative design	project inquiry practices activity teacher tools information community development learners data process environment interaction technology computer social activities analysis online support collaboration group design collaborative	collaboration interaction computer tools information activity community project environment process online activities technology learners social teacher group inquiry practices support development collaborative analysis data design

These were interesting, as it helped demonstrate how important SEL became across the years. Each cluster focuses on interaction, inquiry, and collaboration most strongly. In further sections, more measures looking at this are elucidated to support this finding in the clustering, as well.

Lastly, I also separated the corpus into two decades, 1995-2005 and 2006-2015, to examine whether the most common words presented any interesting differences. While most words were quite similar, a few meaningful differences did appear in the Top 40 word list. This will be discussed in coming sections. Word clouds per decade were created and did not differ drastically, nor did they different drastically with the 20-year word cloud in Figure 1. It was most helpful to look at the entire corpus and the changes in Top 25 words across the decades.

Discussion

While my hypotheses were confirmed by the aforementioned analyses, while further investigation could elucidate even more conclusive findings about these questions. First, a focus on social-emotional learning appears to be increasing moving into the later 2000s. This is evident through the continual increase throughout these years. Words regarding collaboration and a focus on interaction, like "collaborative," "social," and "group" increased on average across the later 10 years. The mean for 2006-2015's use of collaborative is 110.9, whereas 1995-2005's use had a mean of 43.75. Collaboration's mean is also doubled in the 2006-2015 range, social triples, interaction doubles, and support's mean increased by 20. The means for each word were calculated per decade (1995-2005, 2006-2015) to elucidate this comparison. Future research focusing on bigrams or scraping the most utilized words from the entire proceedings, not just the abstracts, could demonstrate this more definitively. The second hypothesis stated centered around technology-based words increasing in the second decade. Words like computer, technology, online, and even design which isn't inherently technology-centric increased in frequency significantly as well. This makes sense as technology became even more prevalent in classrooms after the creation of the internet and its implementation in the later 1990s.

This was interesting, as when the years were split up, "computer" and "online" did not appear at all in the Top 50 words for the second decade. However, words such as "systems," "sciences," and "digital" were more prominent in that time. My conjecture here is that more complex uses of technology were put into use, as opposed to a focus on just Internet or "going online," as is true in the 1996 and 1997 proceedings. This could be confirmed by taking a closer look at words that mean digital, virtual, or computer-related throughout the actual conference papers, not solely the abstracts, by using lemmatization or another natural language processing function in further study.

In a similar vein, out of the Top 20 cosine similarities between years, years within the same decade (2006-2015 vs. 1995-2005) had the strongest connections in 15 out of 20 pairings. This makes sense based on the findings surrounding my first two research questions. Another potential area for future work would be looking at the keywords in titles, in addition to the abstracts. This would allow for further clarity into the contents of the proceedings, as well. This exploration could be done using the same cleaning procedures done here. Because titles have fewer words, but potentially more keywords, this may be even more straight forward and potentially allow for a more granular investigation of these words and trends.

Conclusion

Across the past two decades of conference proceedings, trends have remained similar, yet continued to advance in the Learning Sciences and Collaborative Computing communities. These annual occurrences demonstrate the excitement and ever-changing field of education, especially as technology and a focus on SEL increases. The goal of this project was to elucidate some of these trends and their changes in frequencies throughout the past two decades. This will hopefully inform future conference attendees, researchers interested in entering the field, and practitioners as to what potential next steps, areas of research, or areas of practice could be in a rapidly evolving field. Additionally, similar analyses and graphical representations of inter- and intradecade trends can be utilized to further the understanding of any field, as well as plan for the future of the work.

Figures and tables

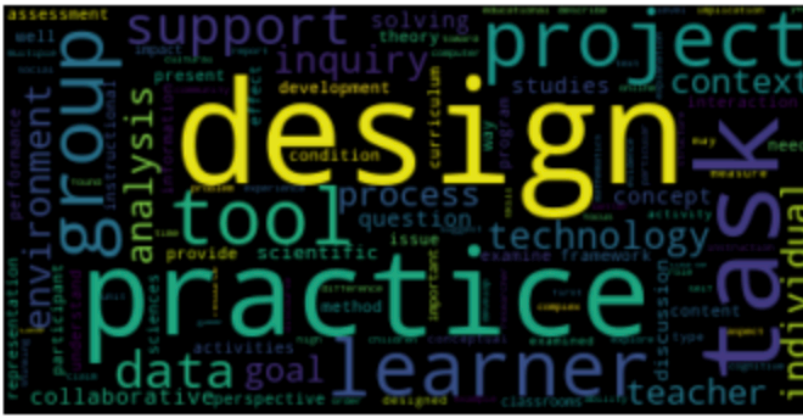


Figure 1: A word cloud of frequencies across both decades

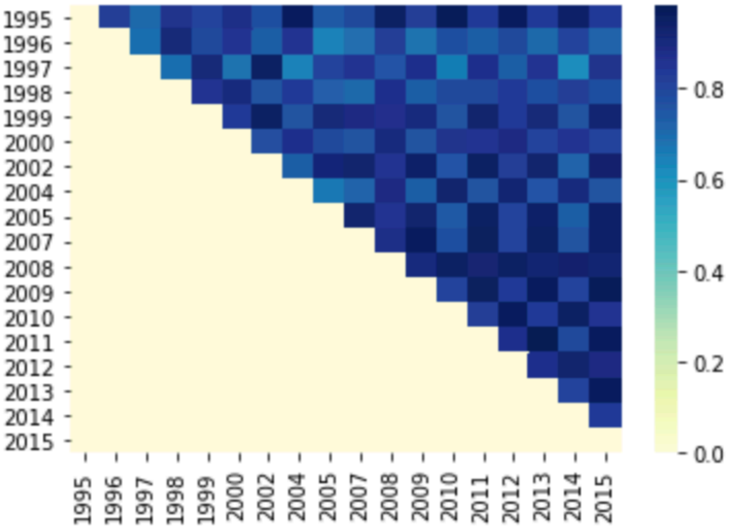
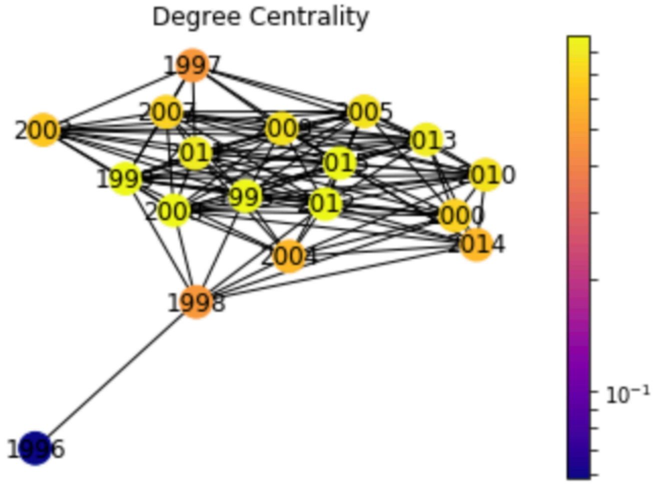
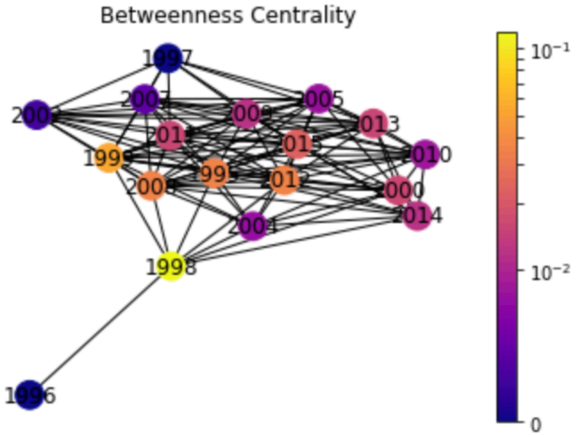
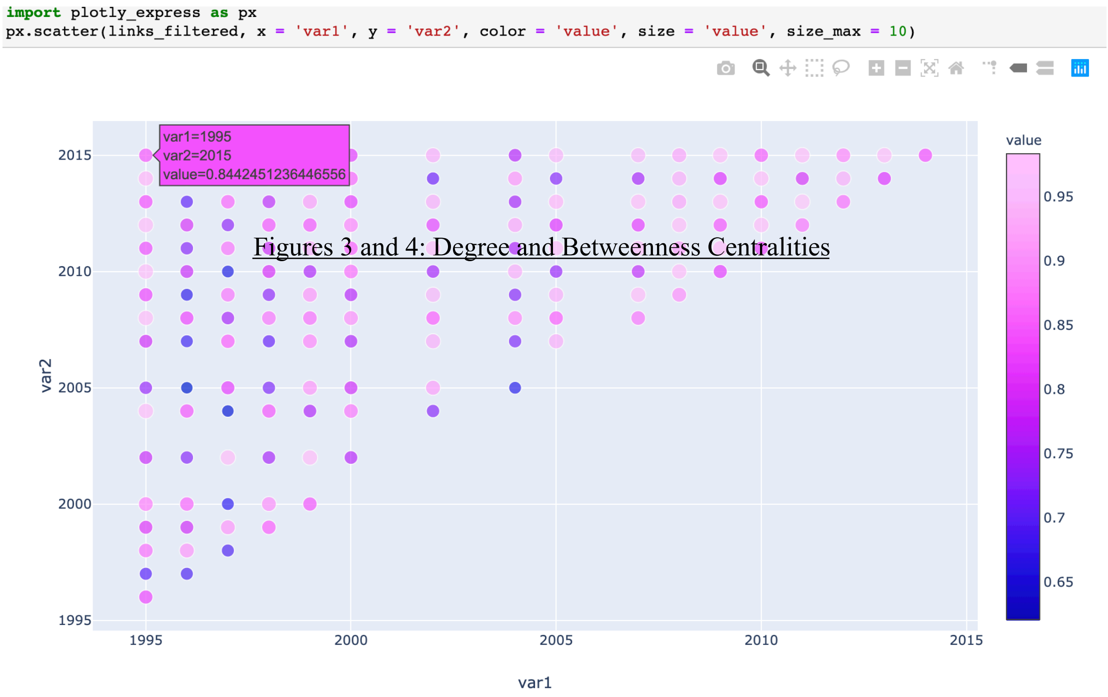


Figure 2: A heat map of cosine similarities





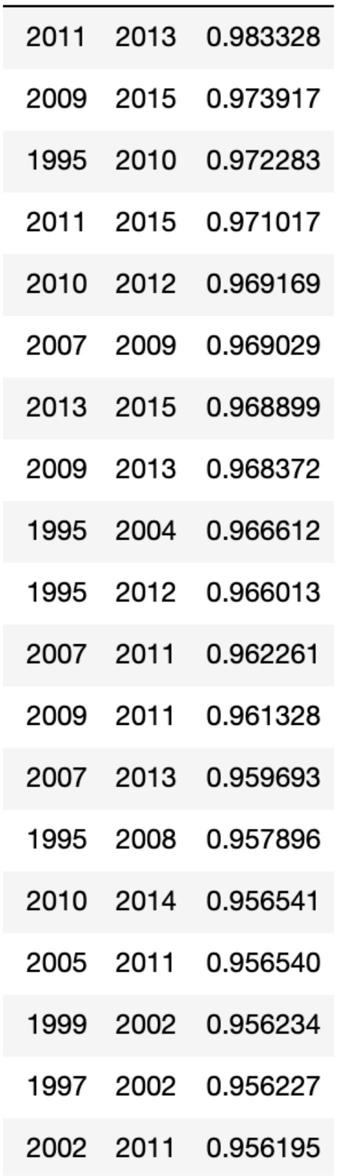


```
import plotly_express as px
px.scatter(links_filtered, x = 'var1', y = 'var2', color = 'value', size = 'value', size_max = 10)
```

Figures 3 and 4: Degree and Betweenness Centralities

Figure 5: Plotly map of year comparisons (interactive)

Table 1: Top 20 most similar vocabularies



2011	2013	0.983328
2009	2015	0.973917
1995	2010	0.972283
2011	2015	0.971017
2010	2012	0.969169
2007	2009	0.969029
2013	2015	0.968899
2009	2013	0.968372
1995	2004	0.966612
1995	2012	0.966013
2007	2011	0.962261
2009	2011	0.961328
2007	2013	0.959693
1995	2008	0.957896
2010	2014	0.956541
2005	2011	0.956540
1999	2002	0.956234
1997	2002	0.956227
2002	2011	0.956195

References

Chen, X., Chen, J., Wu, D., Xie, Y., & Li, J. (2016). Mapping the Research Trends by Co-word Analysis Based on Keywords from Funded Project. Procedia Computer Science,91, 547-555. doi:10.1016/j.procs.2016.07.140

Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014 may 01). Chi 1994-2013. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI 14. doi:10.1145/2556288.2556969

Mccormick, M. P., Cappella, E., O'Connor, E. E., & Mcclowry, S. G. (2015). Social-Emotional Learning and Academic Achievement. AERA Open, 1(3), 233285841560395. doi:10.1177/2332858415603959

Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. Contemporary Educational Psychology, 56, 193-204. doi:10.1016/j.cedpsych.2019.01.009

Ross, B., Chase, A., Robbie, D., Oates, G., & Absalom, Y. (2018). Adaptive quizzes to increase motivation, engagement and learning outcomes in a first year accounting unit. *International Journal of Educational Technology in Higher Education, 15*(1). doi:10.1186/s41239-018-0113-2

Szijarto, B., & Cousins, J. B. (2018). Making Space for Adaptive Learning. American Journal of Evaluation, 109821401878150. doi:10.1177/1098214018781506

Acknowledgments

I would like to extend gratitude to Dr. Bertrand Schneider for his tireless efforts in supporting and teaching our course, as well as William Yao without who served as an exceptional, remarkably hardworking support to the entire course's success. I would also like to thank the Institute of Quantitative Social Sciences at Harvard University for their assistance in working through some of the bugs in this code.