The Grid: A semi-automated tool to support expert-driven modeling

Allegra A. Beal Cohen[1] **Maria Alexeeva**[2]

Keith Alcock[2] **Mihai Surdeanu**[2]

[1] University of Florida, Gainesville, FL, USA
[2] University of Arizona, Tucson, AZ, USA
allegra.ab.cohen@gmail.com

Abstract

When building models of human behavior, we often struggle to find data that capture important factors at the right level of granularity. In these cases, we must rely on expert knowledge to build models. To help partially automate the organization of expert knowledge for modeling, we combine natural language processing (NLP) and machine learning (ML) methods in a tool called the Grid. The Grid helps users organize textual knowledge into clickable cells along two dimensions using iterative, collaborative clustering. We conduct a user study to explore participants' reactions to the Grid, as well as to investigate whether its clustering feature helps participants organize a corpus of expert knowledge. We find that participants using the Grid's clustering feature appeared to work more efficiently than those without it, but written feedback about the clustering was critical. We conclude that the general design of the Grid was positively received and that some of the user challenges can likely be mitigated through the use of LLMs.

1 Introduction

The increasing availability of text data has transformed our ability to model human behavior in social and economic systems. We can now monitor and model phenomena entirely through preexisting text sources like social media, news articles and journal papers. However, these data sometimes fail to capture the causal information we need to build models. For example, news articles may describe what has happened in a region (e.g., "Farmers harvest early") but not why (e.g., "Granivorous birds nearby"). In these cases, one of the best ways to interpret and supplement existing data is to ask local experts for causal explanations of how people think and behave.

Despite the value of expert knowledge, the process of converting it into models remains largely manual and expensive. Fortunately, NLP and ML capabilities have drastically improved since the heyday of expert systems (Devlin, 2018; Ramage et al., 2009; Surdeanu et al., 2022; Schild et al., 2022). If we can partially automate the work required to process expert knowledge, then we can drive more accurate and nuanced modeling of human behavior. While existing NLP and ML methods are powerful, processing expert knowledge presents different challenges than processing large pre-existing text corpora. With this in mind, we combine NLP, ML and visualization methods in a tool designed to satisfy the following criteria based on our experience building models from expert knowledge.

First, NLP tools for processing expert knowledge must allow users to explore text quickly at multiple levels of abstraction. Existing approaches often force a trade-off between digestible summaries and thorough analysis. For example, knowledge graphs can quickly orient users to important topics and relationships, but as the size of the knowledge base grows, topics must be aggregated for the graphs to remain interpretable by humans. Similarly, while Large Language Models (LLMs) are becoming ever more adept at answering questions and providing summaries, they alone do not support multiple levels of abstraction; rather, they require prompts that may be difficult to write during the early stages of analysis when the characteristics and objectives of the user's model are not yet defined.

Second, NLP tools for processing expert knowledge should assign work based on the different capabilities of humans and machines. Many popular topic modeling methods are fully automated, but users are likely to have domain expertise, some familiarity with their corpora, and objectives for analysis and model-building. This expertise should be used to guide the machine. Machines should relieve users of repetitive work and discover patterns that users might not detect, without overriding user decisions. Tools should also support a range of processes and strategies from human users.

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 219–229

In this paper we introduce the Grid (Figure 1), an expert knowledge tool designed to satisfy these two criteria. We first describe the mechanics of the Grid, and then we report results from a user study. Based on our results, we conclude that the Grid supports the efficient organization of expert knowledge and report on challenges and potential solutions for future work on expert-driven modeling tools.

2 The Grid

The Grid is a tool for visualizing and curating expert knowledge. Grids organize textual knowledge into clickable cells along two dimensions. The rows of the Grid represent structural characteristics of the corpus that do not change across topics, and the columns represent topics from the corpus that the user and the Grid work together to discover. The difference between rows and columns is illustrated in Figure 1. Figure 1.a shows a Grid that was created to organize knowledge about the work of an artist, so the rows represent calendar years while the columns represent art media, locations, exhibitions and so on. Figure 1.b shows a Grid that was created to organize interviews with an expert on rice production in Senegal, so the rows represent interviewee and interview date and the columns represent agronomic topics.

The color of each cell in the Grid indicates how much text it contains. Clicking on a Grid cell reveals the sentences it contains and clicking on a sentence reveals the surrounding context (Figure 1b). The user can move and copy sentences between columns, rename columns, and generate columns anchored by keywords. Since the rows in the Grid represent immutable characteristics of the corpus, (e.g., dates or other properties properties of the data points), the user cannot manipulate rows in the same way. The next sections describe how the user and the Grid work together to curate columns through iterative clustering.

2.1 Preparing the corpus

To prepare a corpus for use by the Grid, we first break text into documents. In this paper, our document unit was the sentence. The set of documents is then pre-processed by removing punctuation and stopwords and lemmatizing. Next, the cleaned documents are converted into vector embeddings. For each document, a mean weighted vector is generated using embeddings from the GloVe model (Pennington et al., 2014):

$$V = \frac{\sum_{i=0}^{N} e_i \cdot tfidf_i}{N} \quad (1)$$

where e_i is the vector embedding of word i in the sentence, $tfidf_i$ is its term frequency-inverse document frequency, and N is the number of words in the sentence. Term frequency-inverse document frequency is a statistical method of measuring word importance, where the frequency of a word in a document is compared to how common it is across all documents.

Grids can be *anchored* by specific terms to allow users to focus on subsets of large corpora. For Grids with anchor terms, a subcorpus is generated that contains all documents with the anchor term. This subcorpus is then used to populate the anchored Grid. For example, a Grid anchored by the word "harvest" will contain only documents with the word "harvest" in them, allowing the user to narrow their focus.

2.2 Curating columns

The user and the machine collaborate to cluster document semantic representations or vector embeddings (shortened to "documents" for the remainder of this paper) into columns. This collaboration presents a technical challenge beyond conventional clustering, because user decisions must take precedence over clustering moves made by the machine. We handle collaboration through three types of columns: machine-generated, which contain only the documents clustered by the machine (the first row of Figure 2); frozen columns, created by the user and which the machine is not allowed to change (columns 5-7 in row second row of Figure 2); and seeded columns, which are non-frozen columns that the user has added documents to (see Appendix A for the details on column types).

These three types of columns allow the user to control how the machine contributes to the curation process. The user decides *when* the machine contributes by clicking the "Update" button in the interface. When the Grid updates, all documents outside the frozen columns are re-clustered. The third row of Figure 2 shows the example Grid after the user has requested an update. Note that the user-defined columns written in black text persist, and the machine-generated columns in blue text have changed in response to the user's contributions, highlighting new concepts like "credit."

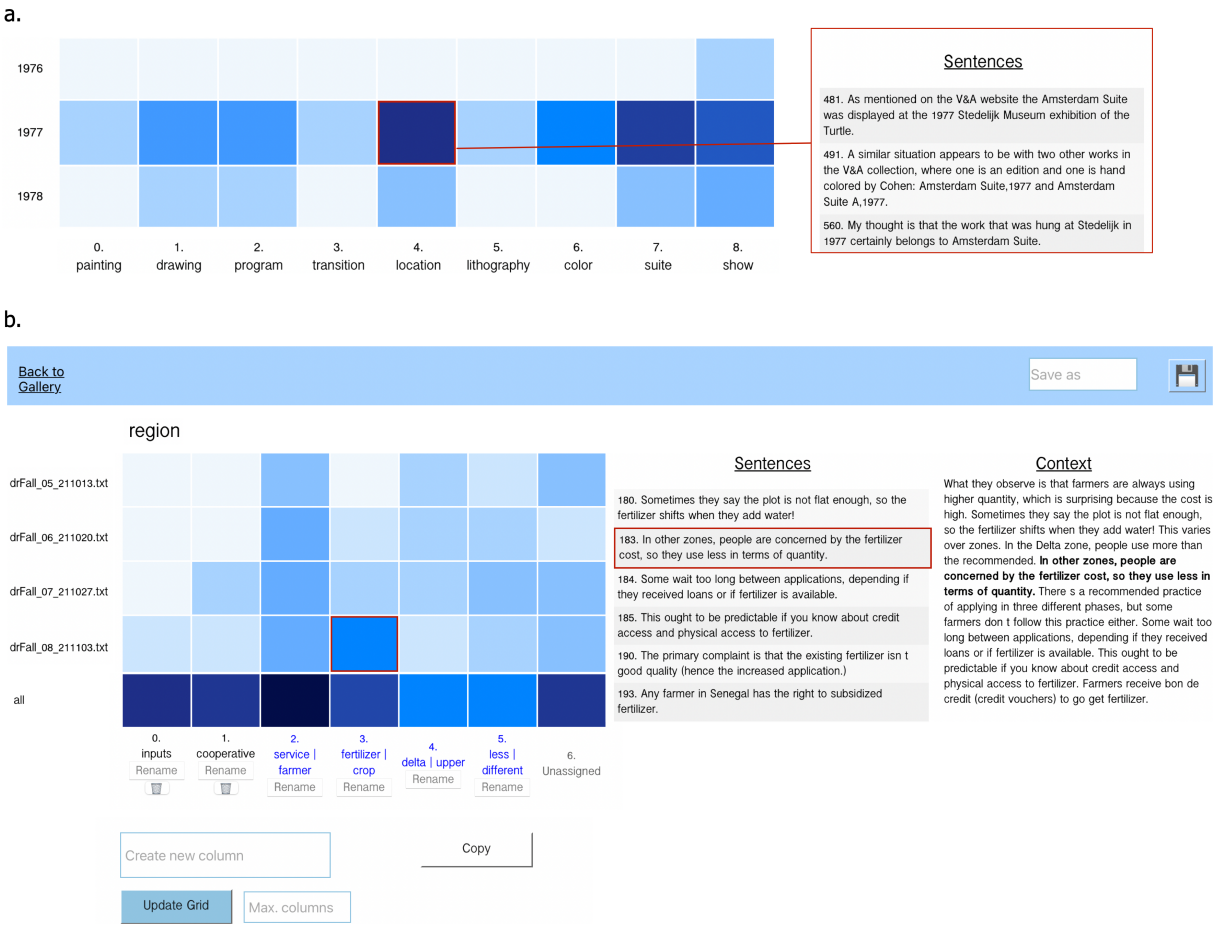

Figure 1: Examples of Grids: **a.** An excerpt of a Grid created using a corpus of emails about an artist, organized along a timeline of when works were made. **b.** An excerpt of a Grid organized by interviewee and date, showing the larger tool interface. Cells in Grids can be clicked on to reveal documents. Documents themselves can be clicked on to show the context, e.g., a sentence in its surrounding interview context.

2.3 Method of clustering

An important feature of the Grid is that documents can appear in multiple columns. To support this, the Grid uses the fuzzy c-means clustering algorithm to assign documents to columns (Bezdek et al., 1984). Fuzzy c-means clustering works by calculating the degree of membership between documents and a given number of k columns. It minimizes the distance between documents and columns, weighted by the degree of membership. Documents are typically assigned random membership coefficients at the beginning of clustering and these coefficients are updated throughout the clustering process. We make one modification to the algorithm: The user-added documents from seeded columns are assigned fixed membership coefficients to ensure that they remain together in the groupings specified by the user.

The number of columns k is selected by running fuzzy c-means clustering multiple times and choosing the k that produces the best model as scored by the Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974). The CH index assigns higher scores to clustering solutions with clusters that contain similar documents internally but that are well-separated from each other. The index is calculated as follows:

$$CH = \frac{(n-k)}{(k-1)} \frac{B}{W} \quad (2)$$

$$B = \sum_{i=1}^{k} n_i dist(centroid_i, meta\_centroid)^2 \quad (3)$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} dist(d_j, centroid_i)^2 \quad (4)$$

where n is the number of documents, k is the number of columns, $centroid_i$ is the average vector embedding of column i, $meta\_centroid$ is the average vector embedding of all documents, B is the

Figure 2: Collaboration on columns between user and machine. Each rectangle is a Grid column, where the color indicates the total number of documents summed over rows. Names in blue indicate machine-generated; names in black indicate user-created.

between distance of the model, and W is the within distance of the model. Frozen columns are included in the CH index calculation because we intend to score the results of collaboration between the user and the machine, not the machine-generated solution alone.

3 Study Methodology

We conducted a user study to explore users' reactions to the Grid and to investigate whether iterative, human-machine clustering helps users organize text more efficiently. We asked study participants to curate an 80-sentence corpus in the agricultural domain (see the section titled *Study corpus*) using the Grid and then take a timed test about concepts in the corpus. Participants were assigned to three conditions with differing levels of automation. In the following section, we discuss the details of the study design.

3.1 Study design

Thirty-nine participants were recruited from multiple domains including development practice, computer science, agricultural engineering and bioengineering. Participants were recruited from academia and included graduate students and faculty members.

We compared the Grid to two versions of itself, resulting in three experimental conditions: treatment, placebo, and control. In the treatment condition, the Grid worked as described in the section titled *Method of Clustering* (Section 2.3). In the placebo condition, the Grid randomly assigned documents to columns instead of clustering them with the previously-described algorithm. The placebo condition was included to test whether participants actually liked the behavior of the Grid or were simply trusting the results of the algorithm regardless of quality (Pan et al., 2007).

In the control condition, participants interacted with a Grid that did no clustering at all. In this condition, participants could create columns using keywords and those columns would be automatically populated, but the machine would not generate any of its own columns. This condition is closest to the spreadsheet-based coding that many social scientists use to process interviews, though it retains the clean visualization of the Grid as well as the automation of keyword-based column creation.

Participants were assigned randomly to the three experimental conditions, with 13 participants in each.

3.2 Procedure

The study was conducted remotely using the Grid hosted on a server. The participants received training for using the Grid, interacted with the tool to

organize the study corpus (the curation stage), and completed a test task and a feedback questionnaire. For more details on the study logistics, see Appendix C.

During the curation stage, each participant was provided with an initial Grid to organize. Those in the treatment condition began with a five-column Grid generated through the algorithm described in Section 2.3. Participants in the placebo condition began with a five-column Grid generated randomly. Participants in the control condition were given a Grid with a single column containing all corpus documents.

3.3 Study corpus

This study used a corpus of expert knowledge about the rice production system in the Senegal River Valley that the authors developed in a related research project. During that project we elicited knowledge from two local experts through qualitative semi-structured interviews. Eighty sentences from these interviews form the corpus for the current user study.

For this study, the rows—the dimension that is associated with structural, topic-independent characteristics of the Grid—represent modeling dynamics since those are commonly used in simulation models. In particular, we manually assigned each document to one of the five modeling dynamics: causes, conditions, decisions, processes, and proportions (see Appendix B).

3.4 Data collection

The study website recorded participants' answers to test questions as well as their written feedback about their experiences with the Grid. Participants were asked to rate their experiences using the Grid on a 5-point Likert scale from "Very poor" to "Excellent." Participants then responded to open answer questions about what they liked and disliked about the Grid, as well as their strategies for using it.

The study website also recorded the actions participants took while using the Grid (i.e., clicks, drags, column creation and updates.) Various summary statistics were calculated from these quantitative data. For example, the amount of work done by participants was calculated as the cumulative number of sentences moved during the curation stage of the experiment. This includes sentences that were moved as part of column creation (e.g., when a user creates a column, we count all the sentences moved by the Grid into that column) as well as dragging sentences between columns. In the placebo and treatment conditions, the cumulative number of sentences moved by the machine during reclustering was also calculated.

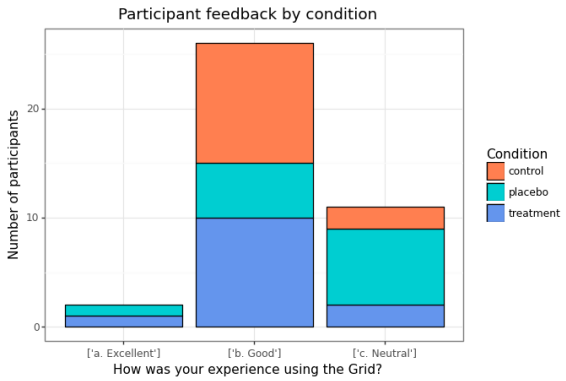

Figure 3: User feedback about the Grid experience. Users were given answer options along a five-point Likert scale, but no responses rated lower than "Neutral."

Participant performance on the test questions was scored by calculating precision and recall. Precision is calculated as the number of answers given correctly divided by the total number of answers given. Recall is calculated as the number of answers given correctly divided by the total number of correct answers (e.g., if a question has two correct answers and the participant gives only one, their recall is 0.5).

4 Results

4.1 Feedback scores

Figure 3 shows the Likert-score feedback given by participants. All participants rated the Grid experience as "Neutral" or higher. Participants in the placebo condition rated the Grid experience as worse more often than participants in the control and treatment conditions. Treating the responses of participants numerically, where 1 = "Very poor" and 5 = "Excellent", the average scores by condition were 3.85 for the control condition, 3.54 for the placebo condition, and 3.92 for the treatment condition.

The feedback in the form of open-ended question responses demonstrated that participants liked the concept and the visualization of the Grid, calling it "easy", "flexible", and "intuitive". For more details on qualitative feedback, see Appendix D.

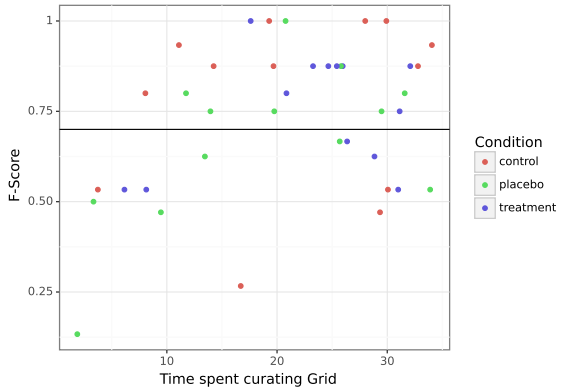

Figure 4: Participant F-scores (y-axis) compared to the time each participant spent curating their Grid (x-axis), colored by experimental condition.

4.2 Test results

The test scores of participants were not significantly different across experimental conditions. The average precision and recall scores were 0.75 and 0.72. We did find that scores (combined into a single F-score for each participant) for participants in the placebo and treatment conditions correlated nonlinearly with time spent building Grids. Figure 4 shows that participants fall roughly into three groups: Those that spent little time curating their Grids and did not do well on the test; those that spent roughly ten minutes or more curating their Grids and did well on the test; and those who spent half an hour or more curating their Grids but did not do well on the test. Figure 4 includes a dividing line at 0.7 demonstrating this rough grouping.

4.3 Cumulative work done by condition

Participants in the control condition moved more sentences on average than participants in the placebo and treatment conditions ($\mu_{\text{control}} = 172$, $\mu_{\text{placebo}} = 85$, $\mu_{\text{treatment}} = 109$; $t(24) = 2.42$, $p < 0.03$ for control-placebo comparison and $t(24) = 1.66$, $p < 0.12$ for control-treatment condition). We do not attribute the difference in sentences moved to the total amount of time that participants spent curating their Grids, because this time was not significantly different between conditions. We also do not suspect that participants in the control condition did more work because they enjoyed using the Grid more than other participants, because participants from the control and treatment conditions gave similar feedback scores. Thus, we suspect that participants in the control condition did more work than participants in the placebo and treatment conditions because the latter were successfully aided by the contributions of the machine.

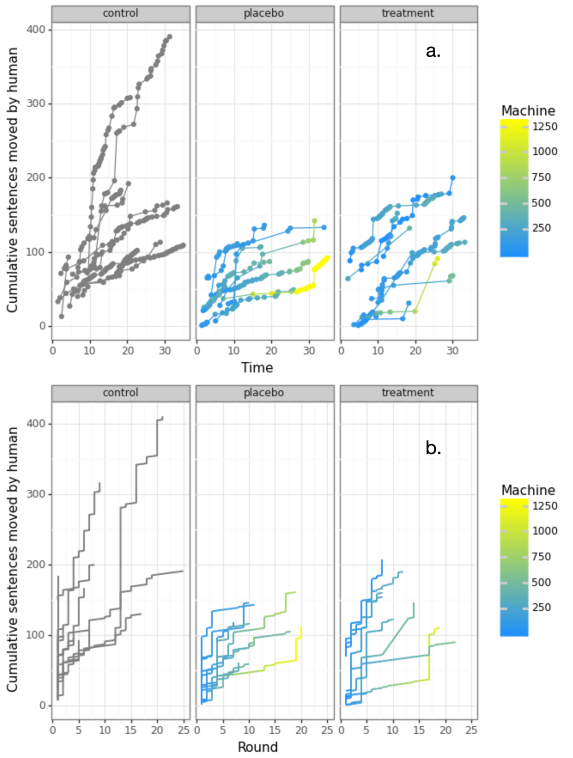

Figure 5: Cumulative sentences moved by the participant and by the machine for each experimental condition.

We examined the strategies that participants used to curate the study corpus. Participants spent time on actions such as sentence dragging and column creation, illustrated in Figure 5a. The y-axis shows the cumulative numbers of sentences moved by individual participants and the x-axis shows time elapsed. Each line represents the activity of an individual participant, and the color corresponds to the cumulative number of sentences moved by the Grid's clustering algorithm. The control condition is plotted in gray because there was no clustering algorithm in that condition.

A variety of user styles is evident in Figure 5a, from an exclusive preference for column creation to progress made almost entirely through sentence dragging. Points that are closely clustered along the y-axis show participants dragging sentences from column to column; larger increases in point elevation indicate that participants are creating columns, i.e., moving a larger number of sentences all at once. The range of strategies shown in Figure 5a is reflected in participants' written feedback. Many reported that the primary benefit of the Grid

was the ability to organize big chunks of information quickly, with some even finding the sentence-dragging feature to be too granular. Others liked that they could move individual sentences by dragging.

4.4 Interaction with the machine

The total number of times participants interacted with their Grids through updating was not significantly different across conditions. However, written opinions about the behavior of the Grid varied.

Participants in both the placebo and treatment conditions reported frustration with the Grid's updating feature (see Appendix D.2 for details). Possible signs of frustration among these participants are visible in Figure 5b. This plot is very similar to Figure 5a, except that the x-axis measures rounds elapsed instead of time elapsed. "Rounds" were counted by how often the user clicks "Update"; for example, Figure 5b shows that most participants across conditions did not update their Grids more than 10 times, while a few updated 20 times. We note possible frustration in the number of updates requested by participants in the placebo and treatment conditions. Several of the lines change color from blue to yellow while maintaining shallow slopes, indicating repeated requests for the machine to do work without corresponding moves made by the participant.

While positive feedback to the updating feature was varied (see Appendix D.2), the data show that some participants worked with the machine rather efficiently. Figure 5b shows a contingent of participants in both the placebo and treatment conditions who accomplished a steadily growing amount of work within ten rounds, perhaps indicating that the machine provided good results in response to participants' first requests. The slope of the lines of these participants is steeper for participants in the treatment condition than for participants in the placebo condition, as we would expect given that the treatment condition was designed to provide better results.

Participants in the control condition were more satisfied with the level of automation in the Grid than participants in the placebo and treatment conditions, even though they did not have access to the column clustering feature. One said, "The coolest feature of the grid is creating new columns and hitting the 'update grid' feature to automatically populate the sentences. It was very cool to be able to parse out a subset of content using key words." Another reported that they "liked the automated aspect of it. Knowing all sentences with the keyword selected would be moved/duplicated to the corresponding column was a helpful way to systematically filter down the information at hand." However, one participant did report that they "did not like that the original column updated on its own based on the remaining information, as it tended to be a bit disjointed."

4.5 Column creation

Participants tended to create between five and ten columns to organize the 80-sentence corpus, with fewer than 20 sentences per column. In the control condition, participants steadily added columns over time, but participants in the placebo and treatment conditions settled on a number of columns within the first ten minutes and then made smaller additions or subtractions. In general, participants in the treatment condition had slightly more columns than participants in the other conditions. Participants in the placebo condition had the fewest number of columns on average.

Participants in all conditions wrote feedback appreciating the automation surrounding keywords and column creation. One participant said that they "liked that the columns included every form of the word rather than just the specific word." A participant in the control condition said that the machine tended to "correctly place information that I thought should be included in [the columns]." One participant reported feeling frustrated that some sentences left over at the end of the curation process did not fit easily into any of the columns they had created.

Participants in all conditions settled on a similar number of columns (the average being ten). However, participants did not all give their columns the same names; the topics in the columns varied more than the number of columns. Table E in the Appendix shows the most and least common words used in column names. The most common words align with the main themes of the interview corpus (e.g., equipment, timing and finances), as judged by the researchers present during the interviews. Participants reported that, during the test, they were able to use the columns they had created to find the relevant information.

4.6 Quality of columns

The quality of Grids is difficult to assess because knowledge curation tasks lack ground truth due to

their inherent subjectivity. However, we can evaluate participants' columns using the same Calinski-Harabasz (CH) index employed in the clustering algorithm (Caliński and Harabasz, 1974). While using the CH index as a measure of quality does tip the scales in favor of the treatment condition, the participants have direction over the clustering algorithm and it is conceivable that human decisions might drive the quality of columns down over time. But when we calculate the CH index for individual participants' Grids over time, we find that 98% of the time, participants in the treatment condition score higher than the highest-scoring participant in the control and placebo conditions. Thus the advantage of using the treatment algorithm persists past Grid initialization.

5 Discussion and Conclusions

The Grid combines NLP, ML and visualization methods to assist users in the organization of expert knowledge corpora. We have presented results from a user study meant to evaluate this combination of methods. Here we draw conclusions about whether the Grid successfully satisfies the criteria laid out in the introduction.

First, we conclude that the Grid allowed users to process the knowledge corpus quickly at multiple levels of abstraction. The organizational power and visualization of the Grid was well-received by participants with diverse expertise and skill sets. Participants in all conditions appreciated the speed with which they could organize information and even participants in the somewhat frustrating placebo condition were able to answer test questions using their Grids. The high test results in all conditions may in part be a ceiling effect; however, we do not discount the role of the Grid in allowing participants to rapidly familiarize themselves with a corpus they had not seen before. Moreover, participants were afforded a large amount of flexibility in how they used the Grid. Participants were able to use different combinations of column creation and sentence dragging to organize information, and they reported preferences for different strategies in the written feedback. Participants often shifted between large organizational moves like column creation and more precise moves like sentence dragging, indicating that the Grid allowed them to work at different levels of abstraction. The number of columns for each participant was similar, but the column names were different, indicating that the Grid allowed participants to organize information in the way that made the best sense to them.

Whether the Grid successfully assigned work based on the different capabilities of humans and machines is less clear. The participants using the Grid's clustering algorithm appeared to work more efficiently than those without it, but written feedback about the clustering was critical. Participants in the placebo and treatment conditions reported that behavior of the Grid's clustering algorithm was confusing and sometimes counterproductive, while participants in the control condition praised the much simpler automated column populating. This indicates that efficiency is not sufficient for a satisfying user experience and that future work on collaborative algorithms should focus on transparency. For example, the inclusion of LLMs in the collaborative process could allow for explanations of why sentences are grouped together in Grid columns.

The Grid tool provides support for organizing expert knowledge in an expert-driven modeling pipeline. While our user study revealed some challenges in the design of such tools, we find the results encouraging and suspect that many of the lessons learned, such as the frustration with the clustering algorithm, may be mitigated in future versions by enlisting LLMs to provide explanations for the user. The Grid can be expanded to include other parts of the knowledge engineering process, such as a semi-automated model generation step after knowledge has been organized. We conclude that semi-automated tools like the Grid can play valuable roles in multiple research communities and have the potential to support more nuanced and local models of human behavior.

The code for the Grid tool is available at `https://github.com/Allegra-Cohen/grid`.

6 Acknowledgments

The authors thank Paul Cohen and Gerrit Hoogenboom for helpful suggestions and support, and the anonymous reviewers for the helpful discussion. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Maria Alexeeva, Keith Alcock, and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed

to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

James C Bezdek, Robert Ehrlich, and William Full. 1984. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, 12(3):801–823.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Erwan Schild, Gautier Durantin, Jean-Charles Lamirel, and Florian Miconi. 2022. Iterative and semi-supervised design of chatbots using interactive clustering. *International Journal of Data Warehousing and Mining (IJDWM)*, 18(2):1–19.

Mihai Surdeanu, John Hungerford, Yee Seng Chan, Jessica MacBride, Benjamin Gyori, Andrew Zupon, Zheng Tang, Haoling Qiu, Bonan Min, Yan Zverev, Caitlin Hilverman, Max Thomas, Walter Andrews, Keith Alcock, Zeyu Zhang, Michael Reynolds, Steven Bethard, Rebecca Sharp, and Egoitz Laparra. 2022. Taxonomy builder: a data-driven and user-centric tool for streamlining taxonomy construction. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 1–10, Seattle, Washington. Association for Computational Linguistics.

A Column Types

Machine-generated columns are those that contain only documents that have been clustered by the machine. Grids are initialized with machine-generated columns. The first row of Figure 2 shows a Grid with five machine-generated columns marked in blue text. The machine names columns by selecting the top two tokens in the column as ranked by tf-idf.

Frozen columns are those that the machine is not allowed to change. Users can create frozen columns using keywords, such as the "labor," "harvest," and "yield" columns in the second row of Figure 2. Each column contains only sentences about its lemmatized keyword. For example, the "labor" column contains only sentences with the word "labor." Users can also freeze existing columns by renaming them (in which case documents in the renamed column needn't contain the user-assigned name). When a column is frozen, the machine is barred from moving documents in and out of it during clustering. Documents in frozen columns also cannot be placed in other columns by the machine, which reduces the amount of organizational work left for the user. Frozen columns are useful when the user has a topic in mind and doesn't want the machine to interfere.

Seeded columns are non-frozen columns to which the user has added one or more documents. When the user drag-and-drops documents into a column, that column becomes seeded (see row four in Figure 2.) During clustering, these user-added documents remain in the seeded column, but the machine is allowed to move other documents in and out of that column. Seeded columns are useful when the user wants to group a handful of documents, but would like the machine to decide which others to include with them.

B Classifying Rows for the Study

The Grid was developed to organize expert knowledge for use in simulation models. Thus, we wanted to organize knowledge into modeling dynamics that bore some resemblance to the code we would write, e.g., conditional language corresponding to `if / else` statements.

We selected five modeling dynamics as rows for Grids in this study: causes, conditions, decisions, processes, and proportions. Documents are classified into rows based on whether they contain information about these dynamics. We define documents as containing causal language if we can identify some X as being responsible for some Y, and containing conditional language if some X is a condition of Y. Documents contain decisions if there is an entity selecting from more than one option. We define documents as containing processes

if there is language about something beginning, ending, or occurring at a specific time or in relation to another process, or if there is language about events occurring in sequence. If some X is compared to some Y, such as with language like "larger" or "more", then the documents contain proportions. Because documents can contain multiple modeling dynamics, we allowed documents to appear in multiple rows. Documents were assigned to rows by hand in this study. Adding automated classifiers is a direction of future research.

C Study logistics

The study was conducted remotely through a website. Participants joined a Zoom room with a researcher present, and then logged into the website using assigned ID numbers. A detailed consent form was provided to which participants agreed in order to continue.

Participants first went through three pages of training, which typically took ten to fifteen minutes, and asked the researcher any questions they had about the Grid. The training was tailored to participants' study conditions.

After completing the training, participants moved on to the next page of the website. On this page, they were given 35 minutes to organize the study corpus using the Grid following these instructions:

> Today you will be working with a corpus of expert knowledge about rice harvesting in the Senegal River Valley. You will have 35 minutes to organize the expert knowledge using the Grid tool. When you are done, you will be tested on the important concepts in this corpus, so please organize your Grid in such a way that you can find information quickly. Think about how you would organize information in your own research; the columns of your Grid should contain what you think are the important themes or variables related to rice harvesting.

During the 35-minute curation phase, when participants in the treatment condition clicked the "Update" button, the Grid returned a new clustering solution using the algorithm described in the *Method of Clustering* section. The Grids of participants in the placebo condition returned random columns. For participants in the control condition, clicking the "Update" button simply removed the sentences from the original column that had already been assigned to participant-created columns. In this condition, the "Update" button helped to tidy up the Grid but did not propose new columns.

After organizing their Grids, participants moved on to a test page that contained their curated Grids and seven multiple choice questions about the content of the corpus. The test questions were designed to strike a balance between broad themes in the corpus and details for which participants would have to read carefully. For example, the first question,

> What could cause a farmer to harvest late? (Select all that apply.)
>
> (a) Bird attacks
> (b) A lack of labor
> (c) Competition for equipment

highlighted the role of labor and equipment in harvest timing (a reoccurring theme throughout the corpus) but also required participants to know that bird attacks cause farmers to harvest early, not late (a more subtle detail in the corpus.)

Participants were given 10 minutes to complete the test using their Grids, at which point they were taken to a feedback page and the end of the study. Finally, participants were debriefed about the condition they were in and the purpose of the study.

D Qualitative Feedback

D.1 Grid concept and visualization

The concept of a tool to quickly organize information into columns was well-received by participants. Participants from all conditions called the Grid "simple," "easy," "convenient," "flexible," "fun," and "intuitive." Participants appreciated the speed at which the Grid allowed them to work and said they liked how it helped them turn disorganized columns into columns that were "well-organized and easier to access." Participants also enjoyed features that allowed them to dig deeper into the Grid content, such as being able to click on sentences to read their surrounding interview context. One participant from the treatment condition said, "It is so flexible ... I can reorganize stuff the way I want ... Super fun to work with."

Participants particularly liked the visualization of the Grid. One said, "I think the visualization with the shading was very intuitive and made the organization process quick and easy to iterate." Participants liked that the colors of the Grid indicated

the distribution of information across columns, saying that it quickly allowed them to infer how "good" their columns were; one participant reported, "I liked the color coding a lot – helped me know which columns were maybe too big, and which were maybe unnecessary or perhaps poorly defined." Another said the Grid was a "good and innovative way to display information to the user."

After completing the study, some participants reported that it had been "fun" and "relaxing." One participant exclaimed, "Where have you been all my life?" and several participants from both the control and treatment conditions signed up to continue using the tool after the study.

D.2 Interaction with the machine

Written feedback showed some frustration among the participants in the conditions that involved clustering. One participant in the placebo condition said, "Very very quirky to use and it was very difficult to get a sense of what the task was." Another participant in the placebo condition reported, "I didn't like how little control I had over what happened during an 'update' – there were different numbers of new columns appearing, etc. I was hesitant to do too many edits once I had a few columns because, again, it seemed like I didn't understand the changes made by the updating." A participant in the treatment condition said, "If I update the Grid, it reorganizes the columns names by itself ... I feel like it is getting out of my hands. The more I want to organize it, more messy it can get."

Positive written feedback about the Grid's updating feature was limited among participants in the placebo and treatment conditions. Only one participant in the treatment condition praised the column clustering, saying "I liked that it would automatically identify and sort motifs."

D.3 Rows and columns usage

Participants in all conditions reported using their column names to navigate to the appropriate sentences based on keywords in the test questions. If the first column they consulted did not have the information needed to answer the question, participants reported that they would move on to the next most relevant column. Most reported that they rarely looked in the rows corresponding to modeling dynamics, but instead used the "all" row that held all of the sentences assigned to a column. A few participants reported that the other rows became useful when the test question was clearly related to modeling dynamics, such as asking what could cause farmers to harvest late.

Participants in all conditions disliked how the study corpus had been organized into rows. Many participants said that they simply did not use the rows because the distinctions between the five modeling dynamics were unclear. In addition, because we allowed a single document to be assigned to multiple rows, participants found that the content of rows overlapped too much. However, others said that a few of the rows were useful, and one participant said that the rows were "practical." In general, participants liked the idea of having rows correspond to modeling dynamics, but found that the actual assignment of sentences to rows was unsuccessful.

E Top most commonly used words in column names

Word	Count
rice	6
machinery	6
labor	6
harvest	5
cooperative	5
harvester	5
loan	4
timing	4
time	4
equipment	4
credit	4
cost	4
farmer	4
season	4