Personalized Language Model Selection through Gamified Elicitation of Contrastive Concept Preferences

Rita Sevastjanova, Hanna Hauptmann, Sebastian Deterding, and Mennatallah El-Assady

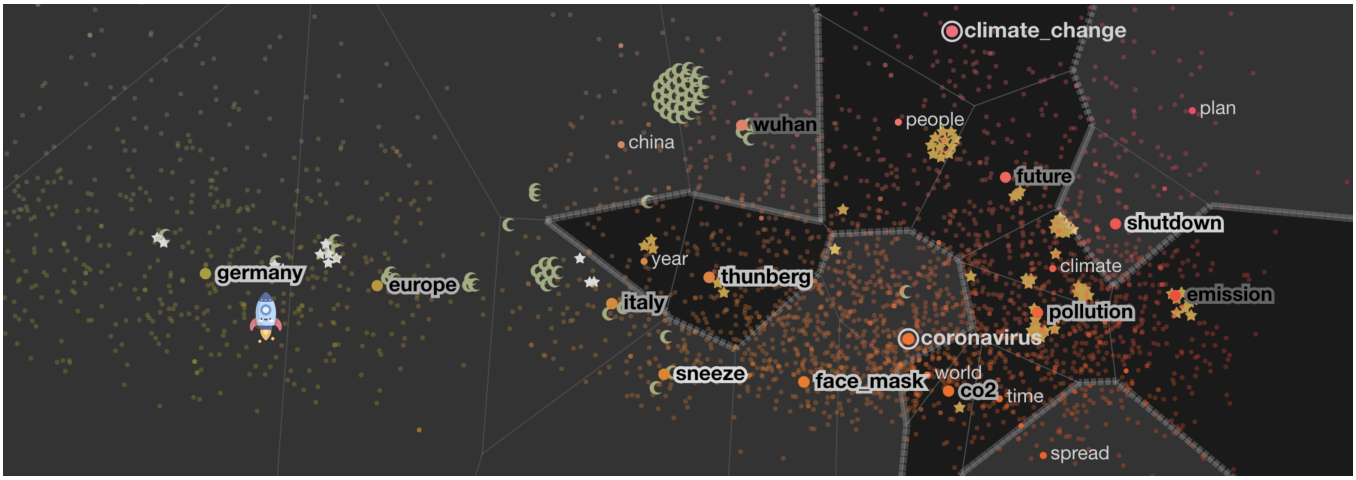


Fig. 1: During the game, users describe two concepts by ten descriptive keywords each. The 2D space of the language model provides visual feedback on the explored language regions. We measure the descriptiveness of the users' input through four quality metrics, which are encoded in the keyword and document design: keyword importance and specificity, document coverage, and precision.

Abstract—Language models are widely used for different Natural Language Processing tasks while suffering from a lack of personalization. Personalization can be achieved by, e.g., fine-tuning the model on training data that is created by the user (e.g., social media posts). Previous work shows that the acquisition of such data can be challenging. Instead of adapting the model's parameters, we thus suggest selecting a model that matches the user's mental model of different thematic concepts in language. In this paper, we attempt to capture such individual language understanding of users. In this process, two challenges have to be considered. First, we need to counteract disengagement since the task of communicating one's language understanding typically encompasses repetitive and time-consuming actions. Second, we need to enable users to externalize their mental models in different contexts, considering that language use changes depending on the environment. In this paper, we integrate methods of gamification into a visual analytics (VA) workflow to engage users in sharing their knowledge within various contexts. In particular, we contribute the design of a gameful VA playground called Concept Universe. During the four-phased game, the users build personalized concept descriptions by explaining given concept names through representative keywords. Based on their performance, the system reacts with constant visual, verbal, and auditory feedback. We evaluate the system in a user study with six participants, showing that users are engaged and provide more specific input when facing a virtual opponent. We use the generated concepts to make personalized language model suggestions.

Index Terms—Language Model Personalization, Gamification, Visual Analytics

1 Introduction

Language models are crucial for many Natural Language Processing (NLP) applications, such as Machine Translation, Part-of-Speech Tagging, or Information Retrieval (IR). Formally, language models are probability distributions over word sequences, and typically generated using statistical or deep-learning-based approaches. Despite their frequent usage, language models still suffer from various issues, such as gender [1] or domain bias [2] stemming from the corpora on which they were trained. One relevant but less frequently discussed issue is that they often lack personalization, as they commonly only depict high-frequency patterns in the training data [3].

While there are methods for personalizing language models to fit user expectations [3], these rely on generating a personalization-appropriate training data set, which has proven challenging [4]. To avoid the tedious generation of training data, we thus suggest a different form of model personalization. There is a variety of models that are adapted to different domain data sets as well as downstream tasks and are made publicly available (e.g., see the Model Hub[1] by HuggingFace or AdapterHub [5]). Thus, we suggest

- *R. Sevastjanova is with University of Konstanz, Germany. E-mail: rita.sevastjanova@uni-konstanz.de*
- *H. Hauptmann is with Utrecht University*
- *S. Deterding is with Imperial College London*
- *M. El-Assady is with ETH, AI Center*

Manuscript received April 19, 2022; revised August 26, 2022.

1. https://huggingface.co/docs/hub/models

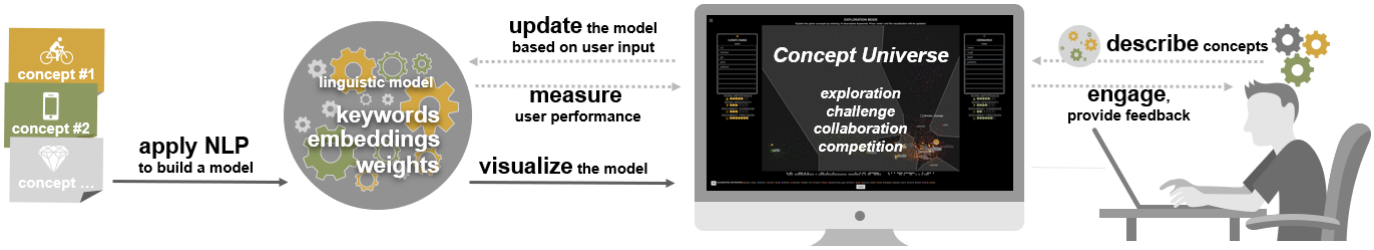


Fig. 2: The analysis process. We use corpora containing documents on multiple language concepts to build an initial linguistic model. Various NLP methods are applied to extract keywords, compute their weights, and extract their embedding vectors. We visualize the linguistic model by applying a dimensionality reduction technique on keyword embedding vectors. The user is asked to explain the given concept names through descriptive keywords. We measure the descriptiveness of their input and provide feedback, motivating them to continue the task.

selecting the model that represents the user's understanding of language from these publicly available models.

To select the best personalized language model, we need to learn how individual users perceive and understand language, also known as their mental model [6]. This commonly depends on various factors, such as a person's cultural, educational, or demographic backgrounds. Learning a user's language understanding is time-consuming [7], no matter whether we elicit it via asking users to describe simple language concepts ("a set of semantically related keywords describing a particular object, phenomenon, or theme" [7]), write text passages, or name word relationships. Thus, any interface that captures their mental models has to motivate them to stay focused on providing high-quality data. At the same time, we need to consider that the context in which we depict the information may influence the quality and the descriptiveness of the gathered data. Hence, to capture these language concepts, we need to either carefully design the targeted environment or simulate multiple contexts and learn an optimal representation from all of them together.

The need for personalizing language models leads us to the following two research questions of this paper: (1) *How can we engage users to share their language understanding and motivate them to provide high-quality data?* (2) *How can we simulate different contexts in which the language is used to learn an optimal language representation and, hence, be able to make a personalized language model selection?*

To support user engagement and simulate different environments in which the language is used, we integrate methods of gamification or gameful design into a VA workflow and present a gameful application called *Concept Universe*. Gamification – the use of game design elements in non-game contexts – has been successfully used for different application domains like crowdsourcing [8], [9], [10], healthcare [11] or teaching [12]. In this paper, we explore the potential of applying gamification in VA for language modeling tasks.

In *Concept Universe* (Fig. 2), we combine multiple NLP methods with visualization techniques to produce an engaging, interactive environment for capturing the users' mental models of particular language concepts. Throughout the seven levels of the game, the users describe two contrastive concepts by representative keywords. We integrate multiple game mechanics to both engage the users and simulate different contexts in which the language is used. To support user engagement, the system stimulates the users to **explore** the language space, to overcome **challenges**, and to **collaborate** with or **compete** against a virtual player. It further provides multi-channel (i.e., visual, verbal, and auditory) feedback on the users' successes at all levels. At the same time, each game mechanics presents a different context in which the language is applied, e.g., a setting with time pressure, or a reflective situation when interacting with a second virtual player. With this paper, we aim at gaining first insights on what type of game elements can be successfully applied in gameful VA processes and how the users perceive this gameful design.

We evaluate our approach through a user study with six participants (linguists and VA experts, and novices in both disciplines). The results show that *Concept Universe* is engaging and motivates users to share their language understanding, independent of their prior gaming experience. We find that concept descriptions are user and (game)context-dependent. We show that not only user-dependent game elements have to be selected for an effective game-experience, but also tailored and gameful visualization techniques. Finally, we use the generated concept descriptions to select the best language model for each user from a subset of models from the AdapterHub [5] repository.

In summary, this work makes the following three contributions to make a personalized language model selection: (1) A detailed design process for integrating gameful design into VA systems; (2) A gameful VA application for learning the users' language understanding through concept descriptions; and (3) An evaluation of the application through a user study with six participants.

2 Related Work

In this section, we provide an overview of methods for modeling language concepts, language model adaptation, and gameful design for VA.

Language Model Adaptation and Evaluation - Language models are probability distributions over word sequences and typically generated using statistical or deep-learning-based approaches. Language models, e.g., transformers [13], are commonly fine-tuned to capture language characteristics for specific domains or tasks, using one of several broad approaches: Domain-adaptive fine-tuning is an unsupervised fine-tuning approach based on masked language modeling task on text from a specific target domain [14]. Intermediate-task training is a model's fine-tuning on labeled data prior to

task-specific fine-tuning [15]. Task-specific fine-tuning deals with adapting a language model to a particular output label distribution [16]. Although used for diverse fine-tuning tasks, these models are rarely used for the personalization [17], [18]. The main challenge for model personalization is the acquisition of the necessary training data [4].

The fine-tuning of language models is effective yet time- and resource consuming. To overcome these limitations, Houlsby et al. [19] introduced *adapters*. Adapters are a lightweight alternative for model fine-tuning, only optimizing a small set of task-specific parameters learned and stored during the adaptation phase, thus, reducing both training time and storage space. AdapterHub framework [5] has brought the advantage of a simple and efficient adapter composition and reuse – one can upload their trained adapters to AdapterHub or HuggingFace repositories and they are available in the framework, supporting the open science practice. Adapters can be trained on masked language modeling as well as specific downstream tasks (e.g., sentiment classification). Adapters have been applied for diverse NLP tasks such as natural language generation [20], machine translation [21], [22], domain adaptation [23], [24], injection of external knowlege [25] language debiasing [26]. Due to the availability of a large number of fine-tuned models (AdapterHub alone has almost 400 adapters), we suggest applying a personalized model selection instead of personalized training, i.e., selecting the best pre-trained or fine-tuned adapter that matches the user's mental model of language. To select the best adapters, we can apply a common evaluation method used on language modeling tasks, i.e., we can analyze the thematic concept separability in the model's generated embedding space [26], [27].

Modeling Language Concepts - For personalization purposes, we thus aim at capturing user interpretation of different thematic language concepts. Ongoing research aims at modeling language concepts through either manual or computational methods. Commonly, crowdsourcing applications are used for this purpose, i.e., for building sentiment lexicon [28], [29], word emotion association lexicon [30], and ontologies [31]. These approaches usually require a time-consuming manual effort. Park et al. [7] have presented a visual analytics system called *ConceptVector* that supports semi-automatic lexicon based concept constructions. In their work, the concepts are generated for a pre-built lexicon, and the system guides the user through the concept construction process using word embeddings. *Topic modeling* algorithms (e.g., LDA [32]) are fully-automatic approaches for concept extraction from a lexicon. Topic modeling is used to extract a set of semantically related keywords found in a document corpus, each set representing one topic (i.e., concept). Despite the usefulness of topic modeling algorithms, the quality of their results depends on the selected parameters and how good they "reflect the characteristics of the analyzed document collection." [33] There has been work done on refining topic modeling results, by applying semi-supervised iterative feedback loop for users to steer the modeling process [34], enabling the users to *vote* on models with a higher quality [35], or through semantic interaction methods within a concept space that is built using a topic modeling algorithm and word embeddings [33]. To capture the user's mental model that is not influenced by pre-computed concept descriptors, we aim at developing an interface where the concepts are created from scratch.

Gameful Design for VA - A manual generation of language concepts can be a time-consuming and tedious task. Therefore, we suggest integrating gameful elements into the VA process to support user motivation. Gamification uses game-based mechanics and aesthetics to engage people in non-game applications. [36] Gamification is defined as "the intentional use of game elements for a gameful experience of non-game tasks and contexts." [37] Game elements, such as points, achievements, leader boards, levels, collections, competitions [38], are used to prompt users to stay focused and motivated while performing a task. According to Ryan and Deci [39], motivation can be twofold. It is either intrinsic (the person is motivated because of the task itself) or extrinsic (the person is motivated due to external factors such as reward). Although the game design has been widely applied in different areas, such as teaching or crowdsourcing (e.g., [8], [9], [10]), it is not yet common in the VA domain. As user motivation plays an important role also in VA processes, recently, we have presented a GamefulVA model [40]. Our model is based on the Knowledge Generation Model (KGM) by Sacha et al. [41], and describes how game elements can support challenging and time-consuming analysis tasks. Due to data overload or complex cognitive tasks, the users of VA applications can loose motivation in each of the KGM steps. Hence, we can measure user interactions to design gameful solutions that motivate the users to explore the data and search for patterns. Also, the verification loop involves complex tasks and actions, such as improving the quality of learning models. In this step, we can apply quality metrics and use them to design engaging game elements (e.g., feedback, development). The knowledge generation loop can benefit from game elements that motivate users in exchanging the gained knowledge. Our design considerations for the integrated game elements (explained in subsection 4.3) stem from the GamefulVA model.

3 Requirement Analysis

The core goal of *Concept Univserse* is to capture the users' understanding of different language concepts. We define the user as a novice individual in language modeling tasks who should be represented by an optimal language model, e.g., in a semantic text analysis setting. For this purpose, we ask the users to describe given concepts with representative keywords. After capturing the users' mental models, we select the best language model based on concept separability in the particular embedding space. As previously described, the two main challenges when designing an interface to capture users' mental models is the potential user disengagement and the impact of the context in which the knowledge is captured. To overcome these challenges, we performed a requirement analysis for both non-functional and functional preconditions that enable an effective extraction of these concept descriptions from different user groups. The following requirements were gathered based on both interviews with VA experts working with language models and related work on data labeling tasks [42].

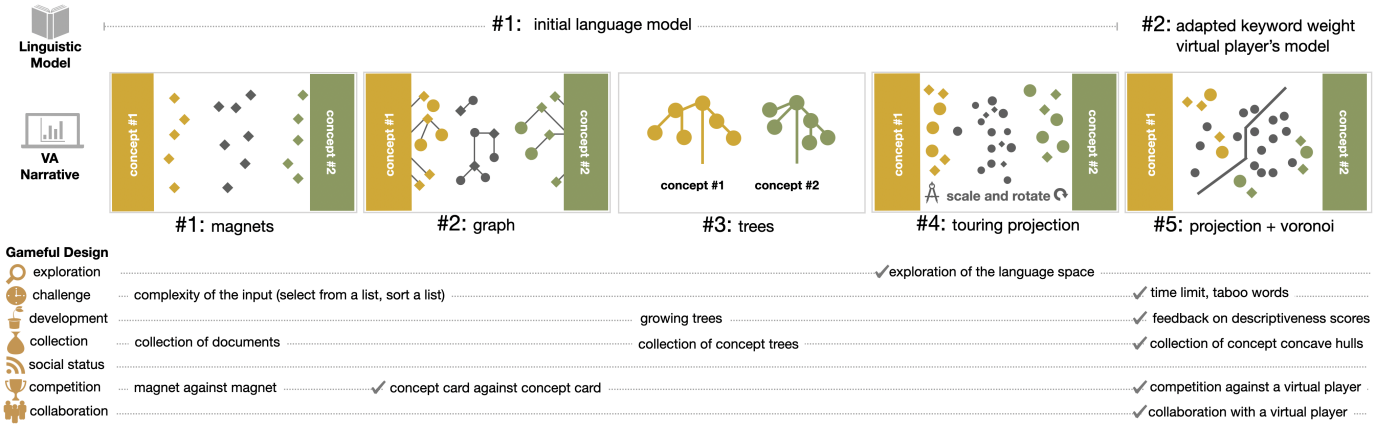


Fig. 3: The *Concept Universe* design required multiple iterations. The linguistic model was designed in two iterations and was settled after the final implementation of the VA narrative. The VA narrative took five iterations; during the design process, we considered multiple visualization techniques for representing the language space, such as a graph, tree, as well as 2D projection layout. The game elements were reviewed according to the particular VA narrative. The checkmark-labeled game elements are implemented in the current design of the interface.

3.1 Non-functional Requirements

We discovered three non-functional requirements, i.e., specifications that are relevant to the system's operational capabilities enhancing its functionality, in particular, for supporting the users' engagement.

NRF1: A gameful VA interface should be **simple and intuitive for different user groups** with varying levels of VA expertise to increase the pool of participants groups. Any background algorithm should be hidden to avoid users getting confused. Also, any feedback that the system provides to the user should be translated into laymen's terms.

NRF2: The system should **integrate a multi-stage analysis process**; each stage should introduce new engaging actions (i.e., game elements) that could be finished within a few minutes. A combination of different stages would make the analysis process more diversified, and hence, we would avoid user disengagement.

NRF3: Errors made during the analysis session **should be editable and recoverable** such that users are not afraid to use the tool and are, therefore, confident in providing their insights and understandings of different language concepts.

3.2 Functional Requirements

We also detected four functional preconditions for the language model algorithm, which need to be satisfied for designing an effective interface.

FR1: We need to create a **basis linguistic model** that presents the expected word distributions for the analyzed concepts. As the goal is to learn an individual's mental model, it should be avoided to use the term "the ground truth" for this default concept representation.

FR2: To create a basis model, one needs a **representative corpus** (e.g., news articles, books, publications) for the concepts that are described during the game. Although the size of the corpus may vary, it should present the underlying concepts appropriately.

FR3: Measuring the **descriptiveness of users' input** is important to provide motivating feedback. This feedback can strengthen users' feelings of success during the analysis process and increase their engagement [40]. To measure descriptiveness, one needs to apply appropriate quality metrics to the input data.

FR4: To help users in performing the task and coming up with new descriptive keywords, one should go beyond a simple feedback visualization that displays the reached scores, i.e., the descriptiveness of their input, and aim at showing a more advanced overview of **keyword relatedness to the optimal model**. This might be achieved by visualizing the underlying linguistic model.

4 Design Process

In this section, we describe our design considerations while developing the *Concept Universe* application, based on the requirements of the previous section. We describe the design iterations for the linguistic model, the visualization, and the game elements. They are described separately since the different concepts required a varying degree of effort. The linguistic model was designed within two iterations, the VA narrative required five iterations, and the game design elements were discussed and adapted based on the discussed VA narrative. Before rejecting a design alternative, it was either discussed in a small focus group of three persons having expertise in VA, or discussed, implemented, and labeled as (un)suitable for the analysis task.

4.1 Linguistic Model

The linguistic model used for processing purposes to measure the user's performance and build virtual players was designed in two design iterations and satisfies the functional requirements **FR1 – FR3**.

4.1.1 Iteration 1: The Initial Linguistic Model

(1) The Linguistic Model (**FR1**) is created from representative corpora. Each corpus contains text documents about one language concept (e.g., Coronavirus, **FR2**). The lexicon includes n-grams (uni-grams, bi-grams, and tri-grams; in the following referred to as keywords) from the corpora that are extracted using the *Document Descriptor Extractor* [43]. The scores, which provide feedback to users on the descriptiveness of their entered keyword, are calculated for each

concept's corpus separately and then normalized to a range between 0 and 1, as explained below.

(2) Keyword Weights: In order to measure the descriptiveness of the users' input (**FR3**), we use a bag-of-words [44] representation and assign each keyword a *keyword importance* score, as well as a *keyword specificity* score. The keyword importance is its average term frequency-inverse document frequency (tf-idf) value [45] in the corpus; i.e., very common keywords in a few documents get a higher weight. The keyword specificity is the keyword's inverse document frequency (idf); i.e., a rare keyword gets assigned a higher weight. If a keyword is not present in the corpora, its normalized weight is assigned to 0. For visualization purposes, each keyword gets assigned an embedding vector using the ConceptNet Numberbatch model[2].

(3) Document Vectors: We use the document vectors to extract a *document coverage* and *document precision* score (**FR3**), which are updated during the gameplay based on the users' input. The document coverage score measures the percentage of documents in the concept's corpus (e.g., documents related to *Coronavirus*) that include at least one input keyword (i.e., a keyword defined by the user during the gameplay) related to that particular concept. On the other hand, the document precision score is determined by dividing the number of documents from the concept's corpus that only contain input keywords related to the specific concept by the total number of documents that contain input keywords from either concept. When the input keywords relate to only one of the concepts, the precision is high.

4.1.2 Iteration 2: Fixing Weights, Modeling Virtual Players

(1) Adapting Keyword Weights: The testing of the initial linguistic model with three linguistic experts showed that the keyword scores, especially the importance score was not representing the intuition of the users. Hence, we adapted the keyword importance score. In particular, we changed the average tf-idf value to the maximum tf-idf value among all documents in the corpus. This means that if a keyword is important in at least one of the documents in the corpus, this keyword is important for the whole corpus (i.e., concept).

(2) Modeling a Virtual Player: Lastly, to engage users to provide concept descriptors of high quality, we integrated virtual players in the system that would either collaborate with or compete against the users. Related work has shown that the social aspect (e.g., playing with or against other users, i.e., social gamification) can influence the user's performance [46]. We thus use the learning-model to confront the users with this virtual player in one of the game levels.

We developed two models for this purpose: (1) one *constant model* that is created based on keyword tf-idf values in our linguistic model, and (2) one *learning model* that is iteratively learned from the user's input. The constant model is created once and contains 100 keywords with the highest importance score, i.e., the maximum tf-idf value [45] in the linguistic model. As a result, the virtual player picks a random word from this model and can achieve optimal performance (i.e., high values of the keyword-descriptiveness scores) during gameplay by utilizing the underlying corpus. Contrary to this constant model, the second model (i.e., learning model) is built on the user's input and gets updated whenever the user inputs a new keyword. The weight assigned to the keyword reflects the frequency with which a user includes a word in the descriptor list. Hence, this virtual player mimics the user's behavior and her mental model of semantic language concepts. To increase the descriptor variability, we use the ConceptNet Numberbatch model to retrieve the five most similar words based on the semantic vectors for each entered keyword, which are additionally incorporated into the learning model.

2. https://github.com/commonsense/conceptnet-numberbatch

4.2 Visual Analytics Narrative

The goal of the visual interface is three-fold: (1) to enable the users to describe concept names by entering representative keywords; (2) to measure users' input using multiple quality metrics and provide feedback on the entered keyword descriptiveness; (3) to display the entered keyword relatedness to the linguistic model. We include two contrastive language concepts at a time for two reasons: (1) it makes the analysis process more diversified, as we can alternate between two concepts; (2) it enables us to apply more game design elements, as the two concept descriptions could potentially *compete* against each other. Each concept is represented by a unique color (**yellow** and **green** accordingly) placed on the opposite side of the screen in a *concept card*. Each concept contains multiple input fields for concept descriptions.

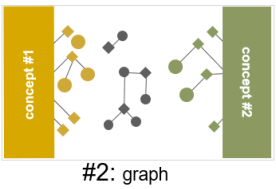
As shown in Figure 3, we had five design iterations for defining visual concepts that would satisfy **FR4**. In the following, we explain the motivation and limitations of each design iteration, whereby iteration 5 (see subsubsection 4.2.5) shows the current design of the system.

4.2.1 Iteration 1: Magnets

#1: magnets

First, we modeled the two concepts as *magnets*. Documents of the corpora were designed as rhombs placed between the two magnets in the middle of the screen. Each entered keyword had an attraction power to the documents in which it was included. The task of the users was to separate the collections of documents according to their concept names. Whenever a new keyword was entered, the positions of the documents in the space were adjusted, i.e., according to the number of occurring keywords in the concept divided by the number of occurring keywords in both concepts. I.e., if the document contained the entered keywords from the underlying concept, they were drawn toward the magnet of this concept. Although the document visualization was straightforward, it lacked information as only an abstract correlation between the concepts and documents was visible.

4.2.2 Iteration 2: Graph



#2: graph

In the second round of design iteration, we aimed to enhance the interpretability of the abstract language space visualization. To achieve this goal, both keywords and documents were illustrated in the area between magnets, using simple icons like circles for

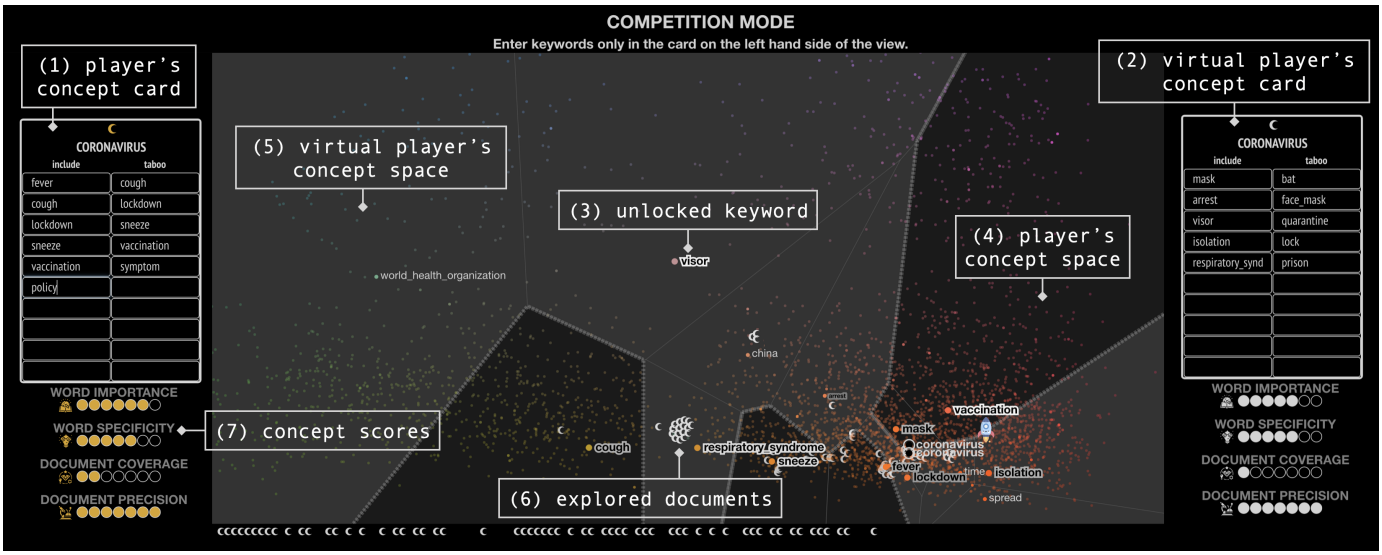
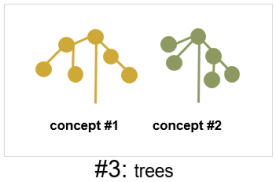


Fig. 4: In the competition level, the user plays against a virtual player. They both have the task of describing one concept, each player on its own. (1) and (2) show the concept cards that the players are filling out. Here, the players are asked to describe two keywords at a time – a keyword to include in the description and a taboo keyword for the opponent. When a keyword is entered, it is visualized in the language space (3). The voronoi areas for both players get updated (4, 5). And finally, we update the position of documents that contain the entered keyword.

keywords and rhombs for documents. A force-directed layout [47] was applied to establish links between documents and their respective keywords. However, while the display of relations between documents and concept cards became more intuitive, the visualization still lacked informative value. This was partly due to the dynamic movement of visual elements caused by the force-directed layout, which resulted in a low level of semantic meaning associated with the document positions.

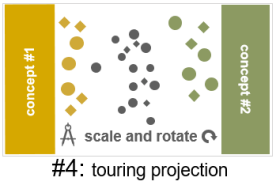
4.2.3 Iteration 3: Trees



#3: trees

In the following iteration, we switched from using graphs to a tree layout. To achieve this, we constructed a hierarchy of the entered keywords for each concept, utilizing word embedding vectors as input for a hierarchical agglomerative clustering algorithm [48]. The tree visualization was updated with each keyword entered by the user, and only the leaf level was displayed to make the tree appear less schematic. In addition, we used a dimensionality reduction method on the language model's keywords to establish a basis for the language space and determine keyword colors [49]. The final design still had flaws; it did not clearly demonstrate the descriptiveness of the users' input or provide clues as to which other keywords may be concealed in the language space.

4.2.4 Iteration 4: Touring Projections



#4: touring projection

Our aim in the fourth iteration was to address the problem of missing spatialization of visualized elements. To achieve this, we utilized the generated concepts from the language model as a reference point to determine the positions of document elements. We accomplished this by extracting keyword embedding vectors and using a dimensionality reduction technique (MDS [50]) to determine their positions on a 2D plane. These positions were then used to create a *spatialized* [51] representation of language concepts. For each document, we determined its position in the 2D space by calculating the average position of its keywords. We proposed a touring projection method that could adjust the projection of keywords and documents based on user input through linear transformation methods such as rotation and scaling [52]. However, we decided against using this method due to its complexity and the difficulty in separating explored concept areas when describing multiple concepts at once. Although this method would have allowed us to visualize differences between users' mental models and the initial language model, it was not practical for our purposes.

4.2.5 Iteration 5: Projection and Voronoi Cells

#5: projection + voronoi

In our final iteration – the current version of the system – we utilize a more user-friendly method to depict the language space. This approach incorporates ideas from our previous design iterations. In particular, we use a dimensionality reduction technique (MDS [50]) to place keywords in a 2D space and use LAB Color Space [53] to assign color codes to keywords based on their positions in this space (also known as keyword semantic coloring [49]). To differentiate the two concepts in the language space, we apply *voronoi tesselation* [54] and create a cell for each entered keyword. We then visually combine cells belonging to the same concept and add borders between cells of opposite concepts. This allows us to highlight the explored regions in the language space for each concept separately. When a keyword is added to the concept card, we increase the size of its icon (i.e., circle) and display its string in the language space to visually

unlock it. Additionally, we update the voronoi borders, since a new voronoi cell is created for the entered keyword, and recalculate the position of each document, which is the average position of the unlocked keywords present in the document. To prevent document overlap, we use a force-directed layout to create edges between keywords and related documents. In cases where the average position of documents is on the voronoi cell that belongs to the opposite concept, we adjust the document positions to the closest relevant keyword to avoid misinterpretation.

4.3 Gameful Design

To satisfy the non-functional requirements, we aimed at incorporating game elements into the VA application to support user motivation and simulate different contexts in which the language is used. Our motivation for the gameful design stems from our previous work [40] that proposes guidelines for an effective integration of game elements into VA processes. In the following, we describe game dynamics that were either discussed or discussed and implemented during the design process.

Exploration "supports user intellectual curiosity" [40]. It is crucial for enabling users to get used to the system, try out the different functionalities without having any negative influence on the analysis results. We integrated this game dynamic as the first level of the game. In the exploration level, they are able to try out different keywords (i.e., enter and delete keywords), explore their impact on the representativeness scores, and learn how to read the visualizations.

Challenge "is a situation in which the outcome requires an effort to accomplish" [40]. During the design phase, we discussed multiple alternatives for integrating a challenge dynamic for the user's engagement. The most simplistic way to integrate this dynamic is by designing the VA process as *game levels* with a varying degree of complexity. The complexity can be specified on different VA concepts, such as the type of users' input (e.g., the task could be to sort given keywords, select descriptive keywords from a list, come up with keywords without any additional help). Furthermore, we can use common challenge mechanics, such as a time-limit that requests the users to perform the task in a limited time frame. In our interface, we integrated the *time-limit* challenge in two stages and implemented them as two game levels. First, we request the users to describe the concepts within a 2-minute frame. Second, we use a time-limit (10 seconds) per keyword. We expect that through the challenge dynamic, the participants would provide more stereotypical input.

Development "shows the evolution of user skills while solving a task" [40]. This dynamic commonly motivates users to continue the task, as they can observe their performance and have an intrinsic desire to improve it. In our application, we provide feedback regarding the entered keyword descriptiveness, and engage users to perform better. We implemented two dynamic development types. (1) During each game level, after a new keyword is entered in the concept card, the system updates the visual representation of the reached keyword descriptiveness scores. (2) After each game level, we provide feedback that summarizes these scores among already played game levels.

Collection "enables the user to gather rewards for performed actions" [40]. We can use different badges, for instance, to show the mastery level of the user according to the achieved descriptiveness scores, or to summarize the created concept descriptions in badge-like representations. We integrate a visual summary feedback that displays all concept descriptions (designed as *concept badges*) gathered during the game.

Social Status "enables sharing user achievements to others with the purpose of social recognition" [40]. We might let the users to share their scores and gathered badges with others. Users describing the same concepts could share their concept descriptions, to learn how unique their descriptions are in comparison to other users. The system is currently designed for a single user; hence, we do not support the social status dynamic yet.

Competition "enables multiple users to compete with each other" [40]. The most simplistic way to implement this dynamic is by letting the users compete against each other while describing the same concepts. Other alternatives include a competition between two concept cards, a competition between a user and the computer (i.e., a virtual player). In our application, the competition is implemented as one of the game levels. In this level, the users compete against a virtual player that makes its decisions based on the learning model trained on the user's input in the preceding game levels. To increase the difficulty, we apply an additional game element called *taboo words*. Taboo words are words that are forbidden to use by the opponent. This game element can motivate the player to develop a strategy for using appropriate words to narrow the performance of the opponent. We expect that through the competition dynamic, the participants would be more reflective, would analyze their own input, and potentially change or adapt it based on the competitor's input.

Collaboration "is known as the efforts of multiple individuals towards one desired outcome" [40]. We discussed several designs to support user engagement through a collaboration dynamic. First, we could let two users collaborate and describe the same concepts simultaneously. However, also a collaboration between a user and the computer can be effective, as the computer might introduce new keywords and inspire the users to look at the concept from another perspective. In our application, the collaboration dynamic is implemented as one of the game levels. In this level, the user collaborates with the computer (a virtual player), while explaining one concept at a time. In the collaboration, the virtual player makes its decisions based on a model that includes a subset of important keywords extracted from the optimal language model. We expect that the collaboration could potentially trigger social thinking, which would motivate the participants to complement the input from the collaborator.

As the effectiveness of the applied game elements is

Game Dynamic	Explanation	Scheme
Exploration	describe two concepts	keywords / keywords
Challenge	describe two concepts in 120 seconds	keywords / time: 120 sec / keywords
Challenge	describe two concepts, you have 10 sec per word	keywords / time: 10 sec per keyword / keywords
Collaboration	together with a virtual player describe concept 1	keywords
Collaboration	together with a virtual player describe concept 2	keywords
Competition	compete against a virtual player, describe concept 1	keywords taboo / keywords taboo
Competition	compete against a virtual player, describe concept 2	keywords taboo / keywords taboo

TABLE 1: The game consists of seven levels. The layout of the interface is adapted according to the played level.

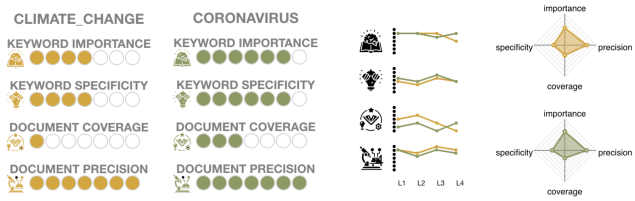


Fig. 5: We implemented two types of the development dynamic: one showing the scores reached within a game level (left), and another displaying the development of scores over multiple game levels (right).

user-dependent (i.e., people have different preferences), we integrated all but the social status dynamic into the interface. With these elements, we aim to engage users while they perform the task. Furthermore, the exploration, challenge, competition, and collaboration dynamics are implemented as separate game levels that enables us to simulate and evaluate different contexts in which the language is used.

5 Designing the Concept Universe

We now introduce *Concept Universe* – a gameful VA system that combines NLP methods (i.e., the linguistic model), visualizations, and game design concepts to provide a gameful playground for learning users' language concepts. The name originates from the gameful design of the interface, representing a *universe* of keywords that get unlocked by the user. The final game provides seven levels, as shown in the Table 1. In the exploration level, the users get familiar with the interface, and in the following (randomized) six levels they are either challenged by a time-limit, they collaborate with, or compete against a virtual player.

To support the system's usability, we integrate several guidance elements in the system. First, we integrate an info-button, that gives a short introduction to the game and summarizes its main visual components. Second, before each game level, we provide a short introduction to the upcoming level, explaining the user's task, and the applied game mechanic (i.e., collaboration, competition). Third, to avoid users' being overwhelmed by the task, we integrate an extra view that displays ten keyword suggestions for each concept, sorted in random order (i.e., a cheat sheet). These are visible when clicking on an eye icon placed on the bottom left corner of the screen. In the exploration view, the suggested keywords are by default visible. We tested this view's influence on users' concept descriptions before finally integrating it into the system. The tests showed that this component is relevant at the exploration level but does not influence users' behavior in the following game levels.

The interface has three main components: two concept cards placed on opposite sides of the screen, and the language space visualization in the middle of the view (see subsection 4.2). The concept cards operate as input fields for users' concept descriptions. Each card gets an assigned color: **yellow** for the left card, and **green** for the card on the right side of the screen. As shown in Figure 4, the concept's name is displayed on top of the card; ten input fields are placed underneath the concept's name. During the game, the task of the user is to fill out these input fields with keywords descriptive for the particular concept. Only at the competition level, the design of the concept card is different. In this level, we introduce a new component called *taboo words*. The objective for the user is to depict the concept using two keywords, both related to the described concept. The first keyword is used as the descriptor as in all other game levels; the second keyword is a forbidden term for the opponent. The forbidden term can be used by the user or the virtual player in her next entry. This should encourage the user to devise innovative tactics (refer to subsection 4.3) to triumph over the virtual player. The taboo word selected by the virtual player is randomly picked from the linguistic model, the same as keywords used as concept descriptors.

To engage the users, we measure the descriptiveness of their concept descriptions and provide visual, verbal, and auditory feedback. The visual and verbal feedback is displayed underneath the concept card and gets updated after a new keyword is entered. The visualization reveals the average descriptiveness scores among the entered keywords. We calculate four descriptiveness scores: keyword importance, keyword specificity, document coverage, and document precision. Each time the descriptiveness scores increase, verbal feedback with a message "Well done!" is temporarily displayed on top of the score visualization. In addition, a short game sound underscores the achievement.

The main visual component is displayed in the center, providing visual feedback on the explored language regions. This component is the fifth design iteration, as explained in subsection 4.2. We use keyword embedding vectors, apply a dimensionality reduction technique (MDS [50]), plot the keywords as small dots in the 2D space, and color them based on their position according to the LAB color space. By default, we display labels for the two concept names and five keywords with the highest average tf-idf weights for each concept. To visually separate the two concepts in the language space, we compute a voronoi cell for each default and entered keyword that are joined for keywords that belong to the same concept. The connected areas present the two concepts: a lighter region represents the first concept, and a darker region – the second concept (in Figure 4). At the beginning of the game, all the documents are displayed underneath the language space grouped according to their concept name. The documents from the first concept's corpus are designed as stars ★; documents from the second

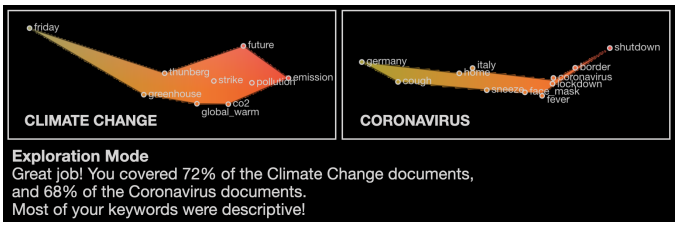


Fig. 6: After finishing the game, the users can explore a collection of *concept hulls* enhanced with a verbal description that reviews the explored regions in the language space and the achieved descriptiveness scores in each game level.

concept's corpus are designed as moons .

Each time a keyword is entered, the language space visualization gets updated. First, auditory feedback signals that the model is currently being updated. For a more game-like experience, a rocket emerges on the bottom of the screen and flies to the position of the entered keyword. The keyword gets unlocked; its representation encodes the importance and specificity scores. In particular, the importance score is encoded in the label's font-size, and the specificity score – in the brightness of its shadow vaccination (i.e., background). Around the unlocked keyword, a new voronoi cell is displayed, and the borders between the two concept areas are updated. Furthermore, the documents that contain the entered keyword get assigned the concept's color, and their position in the space gets updated according to their average keyword score.

The feedback function of game design elements (i.e., scores, score development) can evoke feelings of competence, as it communicates the success of a player's actions [55]. Therefore, we integrate multiple feedback elements in our application. After each game level, a feedback card summarizes the achieved scores both within the played level and among all the played levels so far. We use two representations: (1) a line chart shows the achieved values for each score and each level; (2) a star glyph displays the average score among all played levels. The development of the user's performance is shown in Figure 5. At the end of the game, a summary view gives an overview of the explored regions of the language space and the collection of concept badges gathered during the different game levels. The concept badges are created by utilizing *concave-hulls* [56] placed on top of the explored language space regions, as shown in Figure 6. In addition, a verbal (template-based) summary describes the user's performance. An example is also shown in Figure 6. The goal of this additional feedback is to allow a comparison of the semantic concept models in each level independent of their performance measures.

6 Evaluation of the Gameful Design

The goal of this evaluation is to gain insights into how gameful design can improve the retrieval of language concepts that can then be used to select language models fitting the user's mental model. In particular, we want to gather feedback on integrating gameful design into a visual analytics application and answer the following questions: (1) Is the system engaging, and does it motivate the users to perform the task? (2) Do concept descriptions created within the different game levels have unique characteristics?

6.1 Methodology

In this section, we will describe the methods used during our user study to gather qualitative feedback concerning user engagement and diverse concept descriptions from the different game levels.

Participants: We recruited six participants (three females) with ages ranging from 22 and 29 years, all non-native English speakers. To ensure comparability, we utilize English text data for evaluation, but the method is adaptable to other languages without requiring any modifications to the code. Our study aims to analyze users and game levels comparatively, and as such, we selected participants with similar levels of proficiency in the English language. Two participants (L1 and L2) have a background in linguistics. L1 has experience in working with language models in her research; L2 has experience in annotating word associations. Two participants (V1 and V2) have expertise in VA. The last two participants (S1 and S2) are computer science students, without expertise in linguistics or VA. We are aware that the small number of participants does not allow us to derive statistical results on the system's effectiveness nor the game elements' impact on the participants' performance. Since gamification is not commonly used in a visual analytics setting, in this study, we aim to gain first insights into the potential effects of integrating game elements into a visual analytics approach.

Data: For the evaluation, we created a linguistic model on a news corpus containing two current and widely discussed topics – *Coronavirus* and *Climate Change*. The corpus contains 200 BBC and CNN news articles (100 documents per concept). The news articles about the *Coronavirus* topic have been published between December 2019 and April 2020; the articles about the *Climate Change* topic have been published between January 2019 and March 2020.

Experimental Conditions: Each participant *played* all levels once. None of them was familiar with the interface before the session. The task of the user was to describe the two concepts *Coronavirus* and *Climate Change* while trying to be as descriptive as possible. We tested the system with a few more concepts during the pre-study (e.g., Brexit, Trump's Impeachment, Information Visualization, Visual Analytics), and recognized that due to limited time available for a study, we need to limit the number of concepts. We selected two concepts to be able to compare the results among levels and participants. Every game is started by an *exploration* level, which enables users to try out all the functionalities and get used to the system. In order to get insights about the impact of different game mechanics, i.e., contexts in which the concepts are described, the remaining six levels were ordered randomly. The six levels are: (1) a challenge level that requests users to describe the concepts in two minutes time; (2) a challenge level that requests users to describe the concepts in a limited time – 10 seconds per keyword; (3, 4) two collaboration levels where the user and a virtual player are describing one concept together; (5, 6) two competition levels where the user competes against a virtual player, while explaining the same concept. The competition levels enable the players to define taboo keywords, i.e., words that get locked for the competitor, but can be reused by the player who entered the word.

game mechanic	enjoyment	tension	mental demand	frustration
none	**6.2**	2.6	high	low
challenge 1	5.2	**5.5**	high	high
challenge 2	5.2	**5.4**	high	neutral
collaboration 1	5.3	1.8	low	low
collaboration 2	5.3	2	neutral	very low
competition 1	5.7	3.3	high	neutral
competition 2	**5.8**	3.2	high	neutral

TABLE 2: IMI questionnaire results on the different game elements (range: 1–7; very low–very high).

Procedure: For each participant, we held a two-hour video session (Zoom), which was audio, and screen recorded. We began with a semi-structured interview about the participants' experience in working with language models, and their expectations from a gameful VA system for capturing their language understanding. After introducing the participants to the system, they were able to control the screen and to interact with the interface remotely. We encouraged the participants to think aloud [57] during the analysis session. In each game level, the users were asked to describe the given concepts by entering descriptive keywords. After each game level, the participants were asked to answer six questions from the IMI (Intrinsic Motivation Inventory) [58] questionnaire. After finishing the game, we did another semi-structured interview to get qualitative feedback about the system's design.

Baselines: We define two baselines for our study. To evaluate game elements' impact on produced descriptions, we use the exploration level as a baseline since the users are able to define keywords without being confronted with any explicit engaging elements such as challenge, competition, or collaboration. The second baseline is related to the evaluation of the language model selection task. We use the output generated by the constant model that represents the virtual player (see section 4) as the baseline and compare the concept separability generated by this baseline to those generated by study participants.

6.2 Study Results

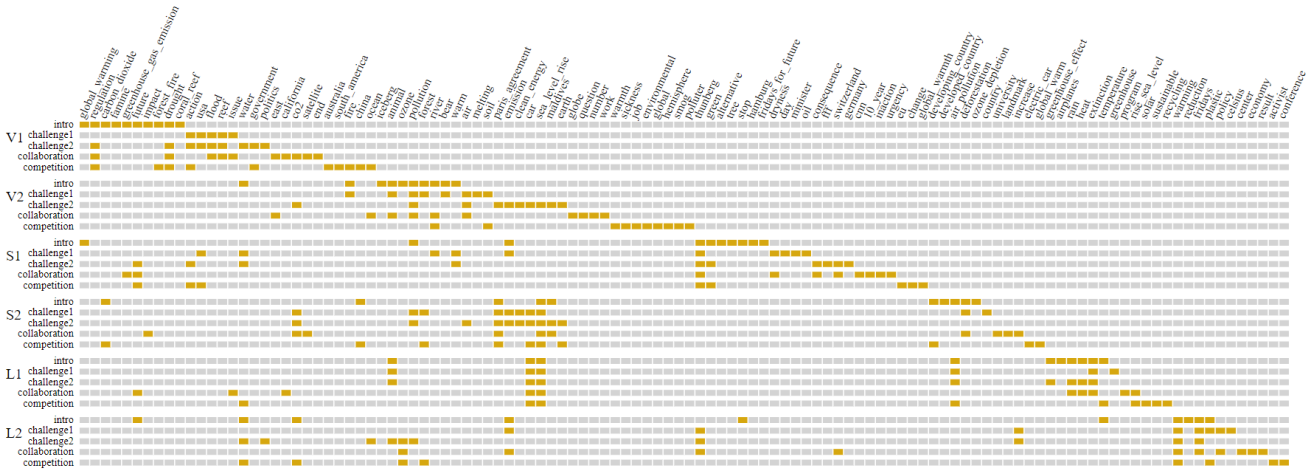
We describe the most relevant topics of our analysis process. We start with the insights for each game level type separately. Then, we summarize the feedback regarding the system's design. The results of the IMI questionnaire are summarized in Table 2.

Exploration Level: All participants started the game with the exploration level in which they were describing both concepts simultaneously by switching between the concept one and two. L1 stated that "*switching between the two concepts is helpful to come up with better keywords.*" Nevertheless, the task was judged as *demanding*, as it was difficult to come up with representative keywords. Despite the complexity of the task, the feedback was positive and participants described the game level to be "*very fun*" [L1, L2, V2, S1]. The design of the interface and the visual and verbal feedback was highly appreciated [L1, L2, V1, V2, S1]. After finishing the first level, V2 concluded that "*It's really fun to play that game!*" Concerning the IMI questionnaire results, the exploration level had the highest *enjoyment* (on average, enjoyment had 6.2 points out of 7) results, and the *challenge* levels the lowest, closely followed by the collaboration levels. We need to consider, though, that this was the first level of the game and the results might show the *wow* effect.

Challenge Levels: All participants described the two challenge levels as mentally highly demanding. We detected two player types among our participants: (1) players who like to overcome challenges (e.g., the time-limit), despite the complexity of the task, and (2) players who don't like to overcome challenges. S1 (one of the participants who prefer challenges) described himself as being nervous before starting the challenge level. Nevertheless, he expected the level to be fun. He stated that "*normally these language concepts are boring for me, but now I just want to get a high score.*" After finishing the game and not filling out all the input fields, he described being frustrated, but he wanted to play this level again, as he believed he could fulfill the task in the given time. The challenge levels (e.g., a time-pressure) lead to some side-effects that are visible in the analysis results. First, due to the time pressure participants felt more frustrated while performing the analysis task, we observed more spelling issues than in other game levels. Second, the time-pressure might motivate people to perform the task more efficiently. In particular, we observed that in the challenge levels, the participants L2, V1, S1, S2 used 91 keywords with an average length of 5, none of them were bi-grams or tri-grams. In the exploration mode, the average word length for these participants was 7. Among the entered keywords were 11 bi-grams or tri-grams. Participants described that they paid less attention to the descriptiveness scores while playing this level. The questionnaire results indicate that in the challenge levels, the participants felt the highest tension (on average, 5.5 out of 7 points), frustration, and mental demand. These levels also had the lowest enjoyment score (on average, 5.2 out of 7 points).

Collaboration Levels: L1, L2, V2, S1 described the collaboration mode as motivating to think of better keywords and emphasized the importance of being better than the virtual player. Before knowing that the interface has a competition level, L1 stated that even though she collaborated with the virtual player, she had the feeling of competition and wanted to reach a higher score than the computer. Participants with a background in VA spent time analyzing the model that was applied for the virtual player. L2 and V2 were assigning roles to the virtual player and themselves. In their opinion, the role of the virtual player was to increase the specificity score, and their role was to think of keywords that cover more documents. L2 stated that "*It was a good combination to reach a high score.*" After playing this level, V1 concluded that he spent more time in analyzing the keywords that were selected by the virtual player than paying attention to the scores. Although the participants thought that these were the most manageable levels, all but L2 described it as the least exciting level in the game. According to the IMI questionnaire, these levels had the lowest frustration scores, and the enjoyment score was 5.3 out of 7 points on average.

Competition Levels: In this level, the players competed against a virtual player in describing the same concept. The virtual player was described as inspiring. The participants paid attention to its input and made their decisions considering it. Participants stated that the virtual player motivated



(a) Keywords created by all participants for the *Climate Change* concept. Taboo words are excluded.

(b) Keywords created by all participants for the *Coronavirus* concept. Taboo words are excluded.

Fig. 7: The study results show that the created concept descriptions are both context and user dependent.

them both to choose descriptive keywords and to think of new keywords that might be related to the descriptive keywords entered by the virtual player. Some unique keyword choices that were influenced by the virtual player (here: VP) were: VP – fahrenheit, L2 – celsius; VP – *coral*, V1 – *Australia*; VP – *flood*, V1 – *ocean*, VP – *impact*, V2 – *cause*, VP – *economy*, S2 – *industry*. L1, V2, and S2 described that the computer's input inspired them to look at the concepts from different perspectives. From the entered ten keywords in the competition level to describe the *Climate Change* topic, V2 entered eight unique keywords that were not used in the preceding game levels (shown in the Figure 7a). L1 stated that in the preceding levels, she was selecting more negative keywords for the *Climate Change* topic. The competition mode inspired her to think of more positive keywords (*sustainability*, *recycling*, *solar energy*). She stated that "*The competition mode motivated me to define the model better. I tried to use more descriptive words. And I did it by looking at the input from the virtual player.*" The participants noticed that the virtual player chose more general words than in the collaboration mode. This observation is interesting since we indeed used two different models for implementing the virtual player for the collaboration and competition mode. For the collaboration, the model contained important words extracted from the initial language model. For the competition mode the model was trained on the user's input learned from the preceding game levels. According to the results from the IMI questionnaire, the competition mode was enjoyed the second most after the exploration mode (on average, 5.7 out of 7 points), although the mental demand and effort were described being high. After finishing all levels, the competition levels were described as the most exciting and enjoyable levels in the game.

Summary on the Role of Game Elements: The different game levels influenced the quality of the concept descriptions created by the study participants. Although these effects are the first observations and are not yet statistically proven, we hope they can help other researchers with designing new gameful VA applications.

In our study, we observed that challenges are mentally demanding; some people like to be challenged, while others are overwhelmed by this game dynamic. Some challenge designs can impact the produced outputs by the users; e.g., time pressure may force people to make more errors and be less extensive in their responses. We observed that the competition is an engaging game dynamic, which can make participants become more creative (see Figure 7). It can, however, also influence the quality of the generated

LM	L1	L2	V1	V2	S1	S2	AA
pre-trained	**0.69**	0.54	0.52	0.50	0.47	0.50	0.45
debiasing	0.44	0.65	0.52	0.42	0.48	0.35	0.45
sst-2	0.62	**0.74**	0.47	**0.55**	**0.63**	**0.57**	0.49
rotten-tomatoes	0.65	0.46	0.52	0.48	0.51	0.50	0.43
imdb	0.64	0.59	0.52	0.48	0.52	0.47	**0.51**
conll2003	0.64	0.48	**0.55**	0.52	0.52	0.47	0.51
average	0.61	0.58	0.51	0.49	0.52	0.48	0.46

TABLE 3: Descriptor separability in the embedding space. We measure the cosine similarity between each concept keyword pair. The concepts are well separated if the similarity between the descriptors within the concept is larger than the similarity to descriptors of the opposite concept. (AA - Automatic Analysis)

outputs, since the user can learn to adapt to the competitor's mental model. Collaboration is sometimes interpreted as a competition; although being less engaging, it can motivate some users to be better than the collaborator.

General Feedback, Design and Usability: All participants provided positive feedback on the design and the usability of the system. Participants liked the variety of integrated feedback channels, especially the sound effects after a new keyword was entered and the verbal feedback that was temporarily displayed each time the keyword descriptiveness scores increased. Participants also positively acknowledged the diversity of the game levels that motivated them to think about the concepts from different perspectives.

Participants emphasized that the gameful design motivated them to perform the task properly. L1 said that "*it is really, really cool. I love it because it is a game! I don't have to just write down the keywords.*" L2, who already had some experience in doing studies for collecting word-association data, described that "*playing the game is so much better than simply writing down the keywords. I have done such studies before, and it is not fun.*" Also, V1 stated that it is a good way to perform a labeling task.

After collecting positive feedback about the gameful design, we aimed at gathering input regarding the effectiveness of the language space visualization, as it was the core visual component in the interface. It turned out that only the two VA experts were using the visualization for the analysis purposes. The novices in VA found the visualization too complex and uninformative. In contrast, V1 and V2 described the visual design as simple and straightforward and used it to get insights about the keyword relatedness. V1 explained that he was using the visualization to "*see what is hidden behind the metrics.*" The language space design was described as effective for the particular analysis task. V2 paid more attention to the clusters that he detected in the visualization to decide which related keywords might increase his keyword descriptiveness scores. During the early phase of the game, V2 speculated "*there must be some patterns in the language space visualization; I guess, I need to figure them out to increase the keyword descriptiveness scores.*"

7 Personalized Model Selection

In the previous section, we showed how the gameful interface motivates users to create high-quality concept descriptions. Capturing user mental models of language enables us to optimize the selection of a representative language model. In particular, since we have obtained the user preferences, we can search for a model that has a matching semantic concept representation in its parameter space (i.e., word embedding vectors). As shown in section 6, game elements have an influence on the generated concept descriptions, e.g., on their quality, length, novelty. Thus, theoretically, the most suitable language model can be selected based on keywords generated in a single game level. This can be useful for applications where the language model should fit to a specific analysis context/circumstances.

In the following, we show how language models can be selected based on semantic concept separability in the model's embedding space. In particular, our goal is to analyze whether different pre-trained and fine-tuned language models produce different embedding spaces for the created concepts. The existence of such differences would endorse the necessity for model personalization. In this evaluation, we use all keywords generated throughout all game levels.

Concept Separability - During the game, the users have chosen keywords that, according to them, are the most descriptive for the given concepts. Our goal is to find an adapted language model (i.e., an adapter) that has a similar representation in its embedding space, i.e., the descriptors of the two concepts should be well separated. In order to measure the concept separability, we first extract context-0 (*decontextualized*) word embeddings for each concept keyword from the pre-trained BERT [3] as well as a random set of adapters that have been fine-tuned on BERT for different downstream tasks, i.e., three sentiment classifiers (sst-2 [4], rotten-tomatoes [5], and imdb [6]), one named entity recognizer (conll2003 [7]), and a language debiasing adapter trained by Lauscher et al. [26] The embeddings are extracted from layer 11 (the layer that captures word semantics [59]) for an input of "[CLS] keyword [SEP]" which is commonly used for language model evaluation purposes (e.g., [26], [60]). We then measure the overlap between the two concepts according to the similarity between their keyword embedding vectors. In particular, we measure the cosine similarity between each descriptor to all other descriptors of the two concepts. The adapter with the largest similarity between descriptors of the same concept (and, hence, the largest distance to the descriptors of the second concept) is chosen as the best representative for the user's mental model of the particular language concepts.

We computed the concept separability for six adapters for each of the six participants and our *virtual player*. The keywords selected by the *virtual player* (i.e., automatic analysis) are the keywords with the highest tf-idf value in the corpus (see section 4), which were representing the *virtual player* during the game. As shown in Table 3, automatic analysis generates keywords with the lowest separability in the embedding space. Although the *Climate Change* concept

3. https://huggingface.co/bert-base-uncased
4. https://adapterhub.ml/adapters/ukp/bert-base-uncased_sentiment_sst-2_pfeiffer/
5. https://adapterhub.ml/adapters/AdapterHub/bert-base-uncased-pf-rotten_tomatoes/
6. https://adapterhub.ml/adapters/AdapterHub/bert-base-uncased-pf-imdb/
7. https://adapterhub.ml/adapters/ukp/bert-base-uncased-ner-pfeiffer/

is described through descriptive words such as *greenhouse gas*, *biodiversity*, *ozone*, etc., it also consists of general expressions such as *region*, *country*, *effect*, and, thus, the descriptors of the *Coronavirus* concept in average are more similar to the descriptors of *Climate Change* concept than to other *Coronavirus* descriptors.

The descriptors representing the two concepts created by the participants on average have a better separability in the embedding space than descriptors created through automatic analysis. Nevertheless, Table 3 shows that the separability depends on the underlying adapter, i.e., there is not one single model that fits all mental models, emphasizing the need for model personalization.

In the following, we provide more detailed insights into descriptor separability for two of the six study participants, i.e., V1 and S1. For this purpose, we use our recent work, i.e., a workspace that enables adapter comparison according to the intersections of their produced embedding spaces to generate explanation visualizations [27]. The adapter with the best separability for V1 is the named entity recognizer conll2003, and with the poorest separability – the sentiment classifier sst-2. However, for S1, the sst-2 adapter performs best and the pre-trained BERT model has the poorest concept separability. The descriptor similarities are displayed in Figure 8a and Figure 8b. Figure 8a shows a PCA projection on embedding vectors for the keywords created by V1. The visualization shows that the sst-2 sentiment classifier is able to group descriptors that are topic-specific, such as *face mask*, *quarantine*, *infection*, *outbreak*. However, conll2003 named entity recognizer has a better separation of named entities related to *Climate Change* (e.g., *greenhouse gas emission*, *carbon dioxide*) as well as geo-locations (e.g., *europe*, *australia*, and *usa*). Contrary to V1, S1 (in Figure 8b) uses very topic-related descriptors for both concepts, thus, the sst-2 adapter has the best separability. These examples show that although the participants described the exact same concepts, their stress lies on different sorts of descriptors. The same goes for the adapted language models; some of them are adapted to capture, e.g., named entities, but others – topic related word similarity. Obviously, different mental models require different types of language models.

8 Discussion

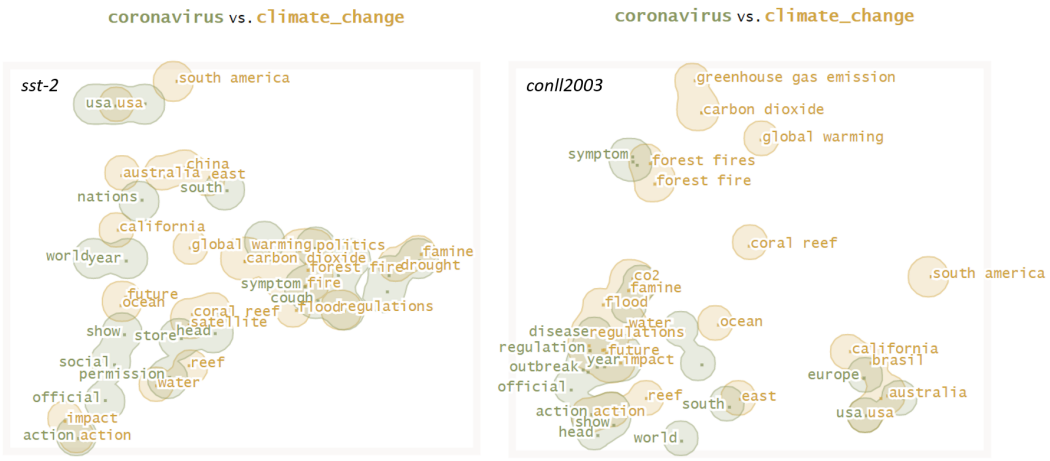
Our study aimed to gather feedback on the integration of gameful design into a visual analytics application and assess whether game elements' have an impact on users' engagement and generated concept description quality. Our study successfully achieved its objective and we obtained first valuable insights into factors to consider when designing gameful visual analytics applications in the future. We identified several key takeaways concerning users' perceptions of gameful design in our application setup. In the following, we will discuss these lessons learned in more detail.

(1) Concept descriptions are user-dependent. Concerning the resulting concept descriptions, we observed differences between users' generated outputs (see Figure 7 and section 7). It confirms the need for capturing the users' mental models to make an optimal language model selection (e.g., AdapterHub [5]) to fit the individual's expectations. For the future, promising research directions are the crowd-sourced collection of language descriptions via games and the large scale simulation of language concept shifts between user groups in different contexts.

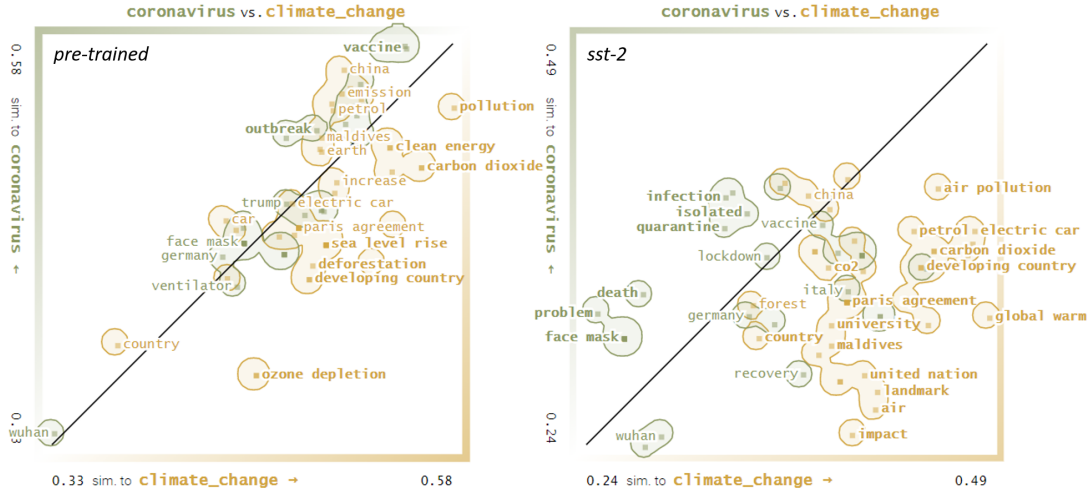
(2) Concept descriptions are context-dependent. Regarding the concept descriptions that resulted from our study, we noticed some variations depending on the different game levels. Specifically, we found that in the more challenging levels, the concept descriptions tended to feature shorter keywords and fewer bi- and tri-grams compared to the exploration level. Moreover, when virtual players were introduced at certain levels, we observed an increase in the novelty and diversity of keywords used, which suggested that users were reflecting on their previous descriptions. However, the changes in the users' concept descriptions depended on the type of virtual player model that was employed. In our experiment, we used two different models for collaboration and competition, respectively. The constant model of the collaboration-player led users to address the virtual player's weaknesses, while the competition-player (i.e., learning model), which was based on past user keywords, encouraged the use of a wider range of keywords. Although our study results are not statistically significant, it is important to note that the *behavior* of virtual players can impact user performance. Their usage should thus be carefully considered in research studies that aim to capture an unbiased mental model of the user. However, in cases where virtual players are built using diverse feature models, they can offer a broad spectrum of potential analysis contexts and be used to steer user behavior toward a desired outcome (interesting for other application scenarios).

(3) Game element preferences are user dependent. In general, the users felt that *Concept Universe* is engaging and fun. They liked the diversity offered by the different levels. While the competition was enjoyed the most by all participants, the preferences for other game elements differed between them. For instance, the time-limit was perceived as pressuring or frustrating by some participants and as positively challenging by others. Again, these results suggest that the usage of game elements in a visual analytics process should be flexible to match the potentially different users' needs. If the engagement aspect is the only focus for integrating game elements into serious analysis processes, future research should provide the user the possibility to select game elements according to their needs or the application should learn user preferences and provide the optimal game element automatically.

(4) Visual design preferences are user dependent. Another diverging preference was reported for our visual design (i.e., the projection with voronoi cells). While VA experts used the visualization to interpret the emerging keyword clusters or keyword positions in the language space visualization, other users instead preferred simple feedback charts. We want to emphasize that both of these user groups are potential target users of our application. The fact that people having more experience in reading visualizations (i.e., a higher level of visual literacy) could use the 2D projection and found it engaging and helpful for specifying new keywords suggests that we need to provide more time for onboarding, i.e., users need time to improve their literacy and get used to the visualization. Moreover, research on the personalization of application designs for different user groups should thus be



(a) PCA projection on embedding vectors. For V1, the poorest separability is achieved by the *sst-2* sentiment classifier; the best separability – by the *conll2003* named entity recognizer. The *conll2003* model particularly well separates named entities such as *cardbon dioxide* or different geo-locations.



(b) Average cosine similarity on embedding vectors between each descriptor and the two concepts. For S1, the poorest separability is achieved by the *pre-trained* BERT model; the best separability – by the *sst-2* sentiment classifier. The *sst-2* model has a good representation for strongly topic-related descriptors such as *face mask* or *air pollution*.

Fig. 8: We use our recent visual analytics workspace [27] to create 2D representations of the concept descriptor embedding spaces. These visualizations help to explore the concept intersection in a 2D (a) as well as high-dimensional space (b).

fostered through methodological aspects (evaluations, discussions), as well as technique-driven approaches in which we measure user performance and derive user preferences.

(5) Generalizing the results. Our system was evaluated with six participants. Although we gained the first insights into the game elements' impact on user decisions and analysis results, a broader study is needed to verify whether our observations are statistically significant. We are currently working on a new version of the interface that integrates the insights gained from this preliminary study. Nevertheless, we hope that this work will motivate more researchers to investigate new approaches toward integrating game elements into visual analytics applications (even beyond the language model personalization task).

Limitations - Playing a game with multiple levels is time-consuming. We need to take care that the game remains engaging, even though requiring some time to be finished successfully. Currently, only two concepts are described within one game session. To make the task more diverse, we could exchange concepts during the game.

The outcome of the game is concept descriptions with a limited size; one could argue whether it is enough information for selecting a language model that is representative not only of the described concepts but that would truly fit the user's mental model of language (i.e., different language concepts). We want to emphasize that this evaluation is only the very first step towards understanding whether we can generalize the information (i.e., concept descriptions/human mental models) that is captured during a single game. The data that is gathered during a game session is relatively small, and we might need to adapt the approach in order to capture a more versatile scope of concept descriptions.

9 Conclusion

We have presented *Concept Universe*, a gameful visual analytics application for capturing users' language understanding through concept descriptions to make an optimal language model selection. We demonstrate our design process,

which includes two iterations for the language model, five iterations for the visualization design, and a parallel iteration concerning the gameful design deliberations. The user study with six participants shows that *Concept Universe* is an engaging VA application that can effectively capture users' concept models. We further make some observations showing that the integrated game elements are able to influence user decisions and performance. We summarize the lessons learned and motivate researchers to consider integrating game elements into analysis processes to strengthen the users' engagement as well as create new analysis contexts (e.g., through a virtual collaborator or competitor) to test how new contexts can steer the users' behavior. More information under: https://concept-universe.lingvis.io/.

ACKNOWLEDGMENTS

This paper was supported by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within projects BU 1806/10-2 "Questions Visualized" of the FOR2111 and the ETH AI Center.

REFERENCES

[1] S. Bordia and S. R. Bowman, "Identifying and reducing gender bias in word-level language models," *arXiv preprint arXiv:1904.03035*, 2019.

[2] A. Fine, A. F. Frank, T. F. Jaeger, and B. Van Durme, "Biases in predicting the human language model," in *Proc. of the 52nd Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 7–12.

[3] H.-Y. Lee, B.-H. Tseng, T.-H. Wen, and Y. Tsao, "Personalizing recurrent-neural-network-based language model by social network," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 519–530, 2016.

[4] M. King and P. Cook, "Evaluating approaches to personalizing language models," in *LREC*, 2020.

[5] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych, "AdapterHub: A framework for adapting transformers," in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 46–54. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.7

[6] J. M. Carroll and J. R. Olson, "Mental models in human-computer interaction," *Handbook of human-computer interaction*, pp. 45–65, 1988.

[7] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist, "ConceptVector: Text visual analytics via interactive lexicon building using word embedding," *IEEE Trans. on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 361–370, 2017.

[8] D. Cirqueira, L. Vinícius, M. Pinheiro, A. J. Junior, F. Lobato, and Á. Santana, "Opinion label: A Gamified crowdsourcing system for sentiment analysis annotation," *ACM WebMedia*, 2017.

[9] N. Venhuizen, K. Evang, V. Basile, and J. Bos, "Gamification for word sense labeling," in *Int. Conf. on Computational Semantics (IWCS 2013)*, 2013.

[10] F. Van Ham and A. Perer, "Search, show context, expand on demand: Supporting large graph exploration with degree-of-interest," vol. 15, no. 6, pp. 953–960, 2009.

[11] R. De Croon, D. Wildemeersch, J. Wille, K. Verbert, and V. V. Abeele, "Gamification and serious games in a healthcare informatics context," in *2018 IEEE Int. Conf. on Healthcare Informatics (ICHI)*. IEEE, 2018, pp. 53–63.

[12] S. Kiesler, R. E. Kraut, K. R. Koedinger, V. Aleven, and B. M. Mclaren, "Gamification in education: What, how, why bother," *Academic exchange quarterly*, vol. 15, no. 2, pp. 1–5, 2011.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[14] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," in *EMNLP*, 2019.

[15] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *ArXiv*, vol. abs/1811.01088, 2018.

[16] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031

[17] K. Bi, Q. Ai, and W. B. Croft, *A Transformer-Based Embedding Model for Personalized Product Search*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1521–1524. [Online]. Available: https://doi.org/10.1145/3397271.3401192

[18] L. Li, Y. Zhang, and L. Chen, "Personalized transformer for explainable recommendation," in *ACL*, 2021.

[19] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Int. Conf. on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[20] Z. Lin, A. Madotto, and P. Fung, "Exploring versatile generative language model via parameter-efficient transfer learning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 441–459. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.41

[21] J. Philip, A. Berard, M. Gallé, and L. Besacier, "Monolingual adapters for zero-shot neural machine translation," in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4465–4470. [Online]. Available: https://aclanthology.org/2020.emnlp-main.361

[22] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, and H. Ney, "Pivot-based transfer learning for neural machine translation between non-English languages," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 866–876. [Online]. Available: https://aclanthology.org/D19-1080

[23] M. Q. Pham, J. M. Crego, F. Yvon, and J. Senellart, "A study of residual adapters for multi-domain neural machine translation," in *Proc. of the Fifth Conf. on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 617–628. [Online]. Available: https://aclanthology.org/2020.wmt-1.72

[24] G. Glavaš, A. Ganesh, and S. Somasundaran, "Training and domain adaptation for supervised text segmentation," in *Proc. of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. Online: Association for Computational Linguistics, Apr. 2021, pp. 110–116. [Online]. Available: https://aclanthology.org/2021.bea-1.11

[25] A. Lauscher, O. Majewska, L. F. R. Ribeiro, I. Gurevych, N. Rozanov, and G. Glavaš, "Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers," in *Proc. of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Online: Association for Computational Linguistics, Nov. 2020, pp. 43–49. [Online]. Available: https://aclanthology.org/2020.deelio-1.5

[26] A. Lauscher, T. Lueken, and G. Glavaš, "Sustainable modular debiasing of language models," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4782–4797. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.411

[27] R. Sevastjanova, E. Cakmak, S. Ravfogel, R. Cotterell, and M. El-Assady, "Visual Comparison of Language Model Adaptation," in *IEEE Trans. on Visualization and Computer Graphics*, 2022 (accepted).

[28] M. Lafourcade, N. Le Brun, and A. Joubert, "Mixing crowdsourcing and graph propagation to build a sentiment lexicon: Feelings are contagious," in *Int. Conf. on Applications of Natural Language to Information Systems*. Springer, 2016, pp. 258–266.

[29] I. San Vicente and X. Saralegi, "Polarity lexicon building: to what extent is the manual effort worth?" in *Proc. of the Tenth Int. Conf. on Language Resources and Evaluation (LREC'16)*, 2016, pp. 938–942.

[30] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–

emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[31] B. Lanser, C. Unger, and P. Cimiano, "Crowdsourcing ontology lexicons," in *Proc. of the Tenth Int. Conf. on Language Resources and Evaluation (LREC'16)*, 2016, pp. 3477–3484.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[33] M. El-Assady, R. Kehlbeck, C. Collins, D. Keim, and O. Deussen, "Semantic concept spaces: Guided topic model refinement using word-embedding projections," *IEEE Trans. on Visualization and Computer Graphics (Proc. IEEE VAST)*, 2019.

[34] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, 2013.

[35] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive learning of topic modeling parameters: A visual analytics framework," *IEEE Trans. on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 382–391, 2017.

[36] K. M. Kapp, *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*, 1st ed. Pfeiffer & Company, 2012.

[37] K. Seaborn and D. I. Fels, "Gamification in theory and action: A survey," *Int. Journal of Human-Computer Studies*, vol. 74, pp. 14–31, 2015.

[38] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From Game Design Elements to Gamefulness: Defining "Gamification"," in *Proc. of the 15th Int. Academic MindTrek Conf.: Envisioning Future Media Environments*. ACM, 2011, pp. 9–15.

[39] R. M. Ryan and E. L. Deci, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist*, vol. 55, no. 1, p. 68, 2000.

[40] R. Sevastjanova, H. Schäfer, J. Bernard, D. Keim, and M. El-Assady, "Shall we play? – Extending the visual analytics design space through gameful design concepts," in *MLUI 2019: Machine Learning from User Interactions for Visualization and Analytics, IEEE VIS 2019 workshop*, 2019.

[41] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Trans. on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[42] A.-L. Kalouli, V. de Paiva, and L. Real, "Correcting Contradictions," in *Computing Natural Language Inference Workshop*, 2017. [Online]. Available: http://aclweb.org/anthology/W17-7205

[43] M. El-Assady, W. Jentner, F. Sperrle, R. Sevastjanova, A. Hautli, M. Butt, and D. Keim, "lingvis.io – A linguistic visual analytics framework," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 13–18.

[44] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.

[45] J. Ramos, "Using TF-IDF to determine word relevance in document queries," *Proc. of the First Instructional Conf. on Machine Learning*, pp. 1–4, 2003.

[46] J. Zhang, Q. Jiang, W. Zhang, L. Kang, P. B. Lowry, and X. Zhang, "Explaining the outcomes of social gamification: A longitudinal field experiment," *Journal of Management Information Systems (JMIS)*, 2023.

[47] F. J. Brandenburg, M. Himsolt, and C. Rohrer, "An experimental comparison of force-directed and randomized graph drawing algorithms," in *Int. Symp. on Graph Drawing*. Springer, 1995, pp. 76–87.

[48] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141–168, 2005.

[49] M. El-Assady, R. Kehlbeck, Y. Metz, U. Schlegel, R. Sevastjanova, F. Sperrle, and T. Spinner, "Semantic color mapping: A pipeline for assigning meaningful colors to text," in *VisGuides Workshop, IEEE*, 2022.

[50] N. Saeed, H. Nam, M. I. U. Haq, and D. B. Muhammad Saqib, "A survey on multidimensional scaling," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–25, 2018.

[51] A. Endert, L. Bradel, and C. North, "Beyond control panels: Direct manipulation for visual analytics," *IEEE computer graphics and applications*, vol. 33, no. 4, pp. 6–13, 2013.

[52] A. Oktaç, "Understanding and visualizing linear transformations," in *Invited Lectures from the 13th Int. Congress on Mathematical Education*. Springer, Cham, 2018, pp. 463–481.

[53] G. Wyszecki, G. Wyszecki, W. Stiles, and J. W. . Sons, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, ser. A Wiley-Interscience publication. Wiley, 1982. [Online]. Available: https://books.google.ch/books?id=HkjM5HNB6jIC

[54] S. Fortune, "A sweepline algorithm for voronoi diagrams," *Algorithmica*, vol. 2, no. 1-4, p. 153, 1987.

[55] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," *Computers in Human Behavior*, vol. 69, pp. 371–380, 2017.

[56] A. Moreira and M. Y. Santos, "Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points," 2007.

[57] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, "Pair analytics: Capturing reasoning processes in collaborative visual analytics," in *2011 44th Hawaii Int. Conf. on System Sciences*. IEEE, 2011, pp. 1–10.

[58] K. S. Ostrow and N. T. Heffernan, "Testing the validity and reliability of intrinsic motivation inventory subscales within assistments," in *Int. Conf. on Artificial Intelligence in Education*. Springer, 2018, pp. 381–394.

[59] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans. of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[60] M. Zhao, P. Dufter, Y. Yaghoobzadeh, and H. Schütze, "Quantifying the Contextualization of Word Representations with Semantic Class Probing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1219–1234. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.109

Rita Sevastjanova is a Research Associate in the Data Analysis and Visualization Research Group at the University of Konstanz, Germany. She received her M.Sc. degree in Information Engineering from the University of Konstanz 2017. Her research focuses on Natural Language Processing, Language Modeling, and the integration of game elements into visual analytics processes to strengthen user motivation.

Hanna Hauptmann is an assistant professor at the Human-Centered Computing Group of Utrecht University, working on intelligent and interactive health systems. She previously worked at the University of Konstanz on human-centered design for interactive intelligent systems, such as explainable AI, personalization, persuasion, guidance, and gamification. She received her PhD at the Technical University of Munich on socio-technical systems for healthy nutrition.

Sebastian Deterding is a Professor of Design Engineering at the Dyson School of Design Engineering at Imperial College London, UK. He received his PhD from Hamburg University, Germany. His research focuses on motivational design and gamification.

Mennatallah El-Assady is a research fellow at the AI Center of ETH Zurich, Switzerland. Prior to that, she was a research associate and doctoral student in the group for Data Analysis and Visualization at the University of Konstanz, Germany, and in the Visualization for Information Analysis lab at the Ontario Tech University, Canada. She works at the intersection of data analysis, visualization, computational linguistics, and explainable artificial intelligence.