H-Matrix: Hierarchical Matrix for Visual Analysis of Cross-Linguistic Features in Large Learner Corpora

Mariana Shimabukuro*
Ontario Tech University

Jessica Zipf†
University of Konstanz

Mennatallah El-Assady‡
University of Konstanz
Ontario Tech University

Christopher Collins§
Ontario Tech University

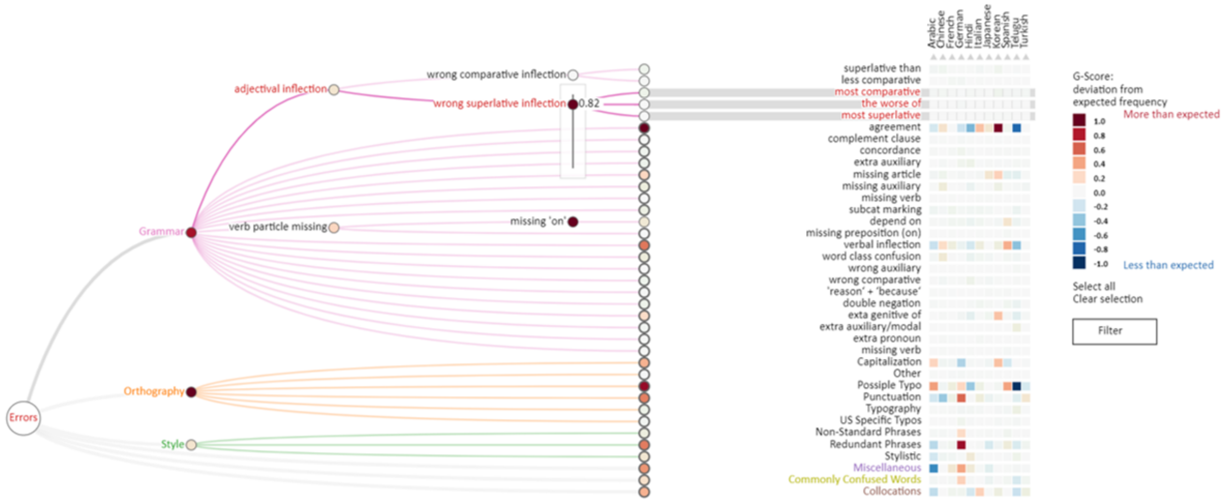


Figure 1: H-Matrix combines a hierarchical view of error categories (left) with a matrix view of errors by language (right). The matrix reveals which errors are more common for each language where the hierarchy summarizes the impact of error categories overall. The matrix rows and the categorical tree branches are highlighted on hover. When hovering on a categorical node, it displays a slider to re-weigh the visual impact of a feature (sub)tree in the matrix.

Abstract

This paper presents a visualization technique for cross-linguistic error analysis in large learner corpora. H-Matrix combines a matrix, which is commonly used by linguists to investigate cross-linguistic patterns, with a tree diagram to aggregate and interactively re-weight the importance of matrix rows to create custom investigative views. Our technique can help experts to perform data operations, such as, feature aggregation, filtering, ordering and language comparison interactively without having to reprocess the data. H-Matrix dynamically links the high-level multi-language overview to the extracted textual examples, and a reading view where linguists can see the detected features in context, confirm and generate hypotheses.

Keywords: Linguistic visualization, learner corpora, matrix

1 Introduction

Second Language Acquisition (SLA) is the research field that studies the process of learning a language other than your native language. When learning a different language, there are language transfer effects (cross-linguistic effects) which can be structural, grammatical or semantic rules from the learner's first language being applied to the second language [14]. For instance, a native Portuguese speaker can misuse the English verbs *lost* and *miss*, because both translate into only one verb in Portuguese, *perder*. We are interested in systematically identifying such transfer effects, and do so using learner corpora, ICLE (International Corpora of Learner of English) [6] and TOEFL11 [3], which are collections of written essays of English learners from different native language backgrounds.

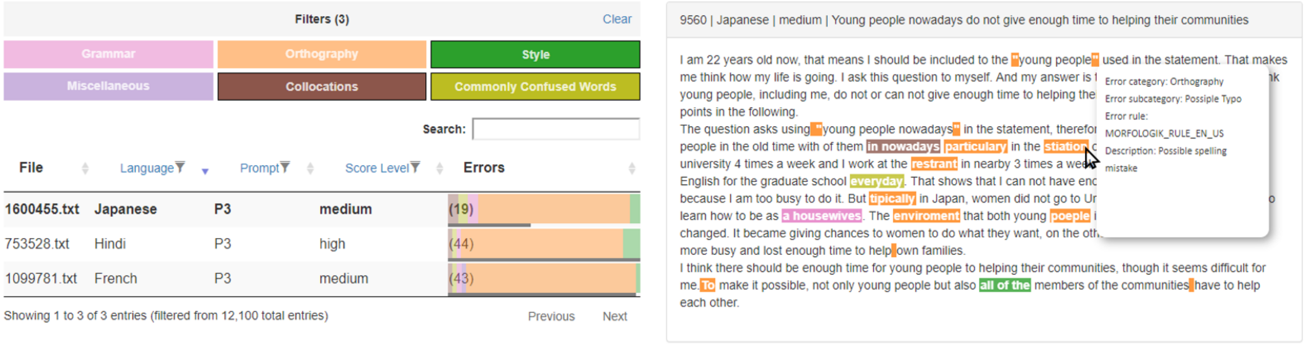
Our user study showed that linguists, and tutors have similar goals towards understanding characteristics one can improve while producing text in a second language. Tutors are focused on a text level analysis to understand performance of learners overtime using smaller sets of essays to tailor their teaching goals. However, linguists are more interested in higher level analysis in linguistic groups, transfer effects and different levels of linguistic features.

Part of a linguist's approach to answer a research question involves processing and extracting a list of features from real-world text, then applying statistical analysis to discover patterns in the data. We can take advantage of existing visualization techniques and tailor them to the specific needs of linguists. Besides analyzing higher level patterns from linguistic features, linguists often rely on having direct access of the text and context of the extracted features.

Our goal is to help researchers interested in transfer effects to identify linguistic patterns in different language groups. The H-Matrix tool allows linguists to perform visual analytics tasks on hierarchical feature sets extracted from large learners datasets. In order to support the different analytics tasks, we designed two views.

The **Hierarchical Matrix View** (Fig. 1) supports high level analysis of a large collection of essays by displaying a matrix with calculated scores based on the observed and expected frequency of features in the corpus. We link the rows of the matrix using hierarchical structure representing the linguistic features. Through the hierarchy, we can perform tasks, such as, aggregation of sub-categories and customization of weights which encode feature importance.

*e-mail: MarianaAkemi.Shimabukuro@uoit.ca
†e-mail: Jessica.Zipf@uni-konstanz.de
‡e-mail: Mennatallah.El-Assady@uni-konstanz.de
§e-mail: Christopher.Collins@uoit.ca



File	Language	Prompt	Score Level	Errors
1600455.txt	Japanese	P3	medium	(19)
753528.txt	Hindi	P3	high	(44)
1099781.txt	French	P3	medium	(43)

Showing 1 to 3 of 3 entries (filtered from 12,100 total entries)

9560 | Japanese | medium | Young people nowadays do not give enough time to helping their communities

I am 22 years old now, that means I should be included to the "young people" used in the statement. That makes
me think how my life is going. I ask this question to myself. And my answer is t nk
young people, including me, do not or can not give enough time to helping the
points in the following.
The question asks using "young people nowadays" in the statement, therefor
people in the old time with of them in nowadays particulary in the stiation c
university 4 times a week and I work at the restrant in nearby 3 times a week
English for the graduate school everyday. That shows that I can not have enc
because I am too busy to do it. But tipically in Japan, women did not go to Ur o
learn how to be as a housewives. The enviroment that both young poeple i
changed. It became giving chances to women to do what they want, on the oth
more busy and lost enough time to help own families.
I think there should be enough time for young people to helping their communities, though it seems difficult for
me. To make it possible, not only young people but also all of the members of the communities have to help
each other.

Error category: Orthography
Error subcategory: Possiple Typo
Error rule:
MORFOLOGIK_RULE_EN_US
Description: Possible spelling
mistake

Figure 2: The essay view combines a table of essays with a panel where essay text is displayed. We can perform filtering and sorting operations when interacting with the table; it also shows overview information about the tagged errors using an in-cell visualization for each row. On the essay panel, the content of a selected essay can be loaded showing its tagged errors. On text hover, we show the error categorization and description.

The **Essay View** (Fig. 2) enables lower level tasks, such as, retrieving specific examples from text, exploring other essays with similar features and validating feature extraction. It displays a table of essays with their meta-information, which allows one to investigate other dimensions of the data (see columns and filters in Fig. 2). A side panel shows the selected essay text with its tagged features.

2 Related Work

Below, we discuss the background on transfer effects in language learning, and visualizations related to our design.

2.1 Transfer Effects

Errors can be used to understand language transfer effects [9, 10] as well as to detect the writers' first language [1, 7]. The former can be seen in Kochmar's work where in compositional distributional semantic models were used to detect errors in word combinations in learner data, more specifically in adjective-noun and verb-object.

Research on Native Language Identification already uses learner corpora to extract features to distinguish between native languages based on English as second language essays [1, 7]. Bestgen et al. showed that using the ICLE dataset we can find common error patterns in English essays that can discriminate native writers of three languages: French, German and Spanish [1].

Errors made by learners are thus realistic and interesting linguistic features related to transfer effect analysis [1, 7, 9, 10]. Similar to the work of Bestgen et al. [1], H-Matrix allows linguists to identify and distinguish features to create linguistic models to deal with specific transfer effects from each first language.

2.2 Hierarchical Matrix Visualization

Matrix based visualizations are widely adopted for linguistic analysis. Mayer et al. used a quadratic matrix (matrix of 4 matrices) to investigate cross-linguistic usage frequency of paired vowels [11]. Sacha et al. combined matrices and line chart to highlight intonation patterns [17]. Keim and Oelke proposed a pixel-based heatmap for literary and authorship analysis [8]. Rohrdantz et al. combined a sunburst to encode hierarchical language families with a heatmap for the linguistic features in a set of outer rings [16]. Similarly, in our work we combine a hierarchy with a matrix in order to enrich the data exploration. However, our hierarchical data is the list of linguistic features instead of the language genealogy.

Hybrid visualization techniques are commonly found when dealing with social networks, multi-variate and high dimensional data, our main data view combines a node-link tree with a matrix visualization. We took as visual and technique references Dendrogramix [2] and Clustergrammer [5]. Dendrogramix combines a dendrogram and a similarity matrix to represent large sets of clusters and data patterns to understand clustering algorithms. Clustergrammer is a heatmap visualization used for hierarchical cluster exploration of high-dimensional biological data, such as, genes and cells.

Lineage [12] and Juniper [13] are visually similar to our approach by combining tree, matrix and table visualizations. However, unlike H-Matrix, which is specifically for analyzing the statistical distribution of hierarchical features, their techniques are to represent attributes of multivariate graphs. Similar to Clustergrammer [5], we designed an interactive heatmap with operations, such as, ordering, filtering and row aggregation to support dynamic changes on the data for analysis. Dendrogramix [2] differs from our design by using the matrix to represent cluster relationships. However, our categories are pre-defined by an expert. Thus, we combined a node-link tree with categories along with a matrix. This combination aims to enrich the exploration using the feature categorical abstraction to perform aggregation and filtering of the rows rather than analyze the hierarchical structure itself.

3 Data Processing

During design and expert evaluation, we used the TOEFL11 as our large learner corpus due to its well balanced sampling methodology.

The **TOEFL11** contains 12,100 English learners' essays submitted to the TOEFL language proficiency test [3]. It has 11 L1's (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish), along with information on the score level (low, medium and advanced). They were also evenly sampled from 8 retired TOEFL independent prompts (topics).

For the **feature extraction**, we used the LanguageTool API which is an open source proofreading software for English and other languages [15]. This API uses a rule-based model to extract errors. The rules are also maintained by the community. It offers a broad categorization of these rules used as base for the prototype. The rule categorization was later on customized by our expert collaborators. However, this paper describes the final and improved version.

4 H-Matrix Visualization Design

H-Matrix contains three main coordinated components for the matrix, hierarchy, and essays; each is described below.

4.1 Matrix Visualization Component

In order to support large scale statistical analysis of the frequency of the extracted features, we built a matrix visualization where the rows are the linguistic features (error rules) and the columns represent the learner data grouped by first languages (right side in Fig. 1). A matrix visualization is well known and adopted by linguists when performing visual analysis [11,16,17]. Therefore, we took advantage of it by having the main component as a familiar tool to our experts.

For the **encoding values** in the matrix cells, we have 2 options: raw frequency and the g^2-score. The raw frequency is the actual number of times a feature or an aggregated set of features (tree leaf) occurs in the dataset per first language. The g^2-score, or Dunning Log Likelihood value, is an expectation measure based on the frequency of the occurrence of a feature in a language group relative to the whole dataset [4]. The g^2-score will be further from zero the more the frequency differs from an expected uniform distribution. For example, a feature observed 100 times within a dataset of 10 first languages, it is expected to appear 10 times in each language. When a language's actual frequency is much lower or higher than 10 it will have a high g^2-score, indicating the deviation from expectancy. We indicate positive or negative scores for cells producing more or less errors than expected, respectively.

For **normalizing the displayed values**, we have 4 options: global, row, column, and none (no normalization). We anticipated that global normalization can help with an overview analysis of the corpora. Row normalization can allow analysis across language groups per feature. The column option helps for analysis of features within a language. Naturally, we also have an option of display the raw frequencies in case one is interested in absolute values.

To facilitate the visual analysis for different cases, we created two options for **color scales**: linear and divergent. The linear scale can be applied to both types of scores, however, the divergent scale is only available for the g^2-score as it can be positive or negative.

Another common operation for working with a matrix is **sorting**. H-Matrix supports sorting the rows by their summed score, or by a selected column. One can also sort the columns by their summed scores. These operations support tasks, such as finding high or low aggregated scores per feature or language.

Each cell in the matrix provides a tooltip pop-up with extra information on its values. We provide the raw and the calculated measures. Cells, rows and columns can also be selected in order to filter the **essay level view** with essays containing the selection of errors by clicking in the *Filter* button.

In Fig. 1, the subcategory *agreement* has the highest aggregated g^2-score (darkest node in the *Grammar* subtree). The *agreement* cell color in the *Korean* column shows that native Korean learners produce more agreement errors in English than the other languages. This English transfer effect happens because the Korean language does not have agreement rules, such as, verb and subject agreement.

4.2 Tree Component – Hierarchical Features

In order to perform linguistic analysis, an expert can extract different features from different classes to understand multiple facets of a text. Therefore, experts usually work with different sets of extracted features which can be grouped together in a personalized hierarchical structure. In this paper, we worked with linguists to label and categorize the extracted errors from essays into 6 main categories: grammar, orthography, style, commonly confused words, collocations and miscellaneous (left side in Fig. 1). Naturally, this categorization is arbitrary, and can vary depending on how the features were extracted, and the expert's affinities and interests.

We decided to implement aggregation and categorization on the low level features due to large scale of the data which this tool intends to support. For broader categories, such as, *grammar*, they can have sub-categories and sub-sub-categories containing the rule-level features. For instance, the rule "many kinds of," which signals the plural disagreement of the word *many* and the followed noun, is under the classification "grammar/agreement/nominal agreement (plural)/many kinds of."

Through **interaction** we enable two main operations on demand: dynamic aggregation and re-weighting of (sub)categories nodes. These operations allow experts to change the granularity and importance of the data on-demand.

Without needing to go back into the data processing step, one can investigate different levels of abstraction regarding the error rules being analyzed when **collapsing or expanding nodes** with a left click. Right click opens a menu that offers options to expand/collapse all levels in a node. When a group of nodes is collapsed all the features under it are aggregated. Dynamically, these changes will reflect in the matrix cells normalization and color scale accordingly.

When hovering on the nodes, a slider component is shown. This slider enables **dynamic changes in weight** for all the features under the selected tree node. This slider can be used to reduce or neutralize branches or leaves of the tree, therefore, it removes unimportant features on-demand depending on the data analyst's needs. When the weights are changed the appropriate weight distribution is reflected to siblings and children of the node, all the way to the matrix cells scores, depending on the color scale and normalization used.

4.3 Essay Level Component

This view is initially populated by all the essays in the dataset. However, the hierarchical matrix visualization view can also be used to filter the essays according to their language and/or type of error feature. The essays are shown in a table which can be used to load a specific essay in the text view (Fig. 2).

We display a table where each row is an essay, and the columns can vary based on the dataset's meta-information. In the case of the TOEFL, the columns are file name, language, prompt (similar to topic, is the question learners were prompted when writing the test), proficiency level and the amount of errors. The information in these columns is straight forward, except the errors column.

The data value in the errors column is the number of errors contained in the essay, while the cell contains a visual representation of the number of errors and the proportion of each error category contained in the essay. We render 2 horizontal bar charts, the top bar represents the distribution of the errors in each of the categories; the lower bar encodes the number of errors which is also displayed in for of text. Using this **in-cell visualization**, one can make visual comparisons of the different error categories and their proportions included in each of the essays. The bar encoding the number of errors helps to visually compare the amount of errors in one essay versus the whole selection listed.

The essays table allows **sort** to be performed in all the columns. However, **filter** can be applied only in the language, prompt (topic) and proficiency columns. Along with the table, this view also has an error category filtering panel. The **category filtering** options when used are applied to the essay table facilitating the exploration of essays containing a specific error category.

When hovering on an essay row, a **tooltip** pops up with a preview of tagged errors from the essay. The tooltip lists each of the tagged errors in a snippet with surrounding context words. If a row is clicked, the **full text** of the essay is displayed on the text panel along with their tagged errors. When hovering on the tagged error, more information, such as, error category, sub-categories if applicable and the rule is shown. This view can be used to verify hypotheses about the data exploration, and to validate the feature extraction step.

5 Expert Case Studies

We designed a semi-structured interview combined with an exploratory analysis session, where participants had hands-on experience with the tool. They were asked to describe the methods and processes they would use if they were to identify error patterns in a large learner corpora, as well as to describe their expectations for a tool that would support their processes.

Participants We ran the study with 6 participants. The experts who volunteered for the study are computational linguists, linguists and language tutors. All had relevant experience with language and second language acquisition. The experts fields of interest included: writing support, pedagogy, psycho-linguistics, regressive transfer effects, early language acquisition and bilingualism.

Exploration Session Following the interview, presentation of H-Matrix functions and scheme of the TOEFL11, participants performed a pair analytic session where they had full control of the tool. Meanwhile, the investigators were available to clarify any of the system functionalities, and if needed, to suggest interesting tasks to the participant, *e.g.*, find the highest occurred grammatical error for a specific language and seeking text examples of it.

Results The **Language tutors** described their work to be focused on personal or small groups of students. They also explicitly stated their pedagogic stance to not point out errors, but rather motivate learners to produce *anything* in their second language. Cohesion and vagueness of arguments were also interests. They seemed overwhelmed when presented with H-Matrix and showed resistance to most of the hierarchical matrix functions. However, they showed interest in the text view for displaying the tagged errors, but suggested some changes in order to make the view more useful for them.

Some of the suggested changes were to add another text panel for comparing two different essays side-to-side, to add more detailed explanations for the tagged errors and functions for editing, grading and commenting each essay as a tutor. They also suggested a mechanism similar to a file version manager to track student development.

Regarding the hierarchical matrix view, most of their suggestions were related to removing functions, such as, re-ordering of the rows, different levels for the error categories (they preferred high-level only), weighting of the nodes, and the g^2 score. They also would want to work with much smaller groups, if not only individual essays.

In general, they were confused with the interactions and scale of the data presented. For instance, they think that the granularity and categorization of the hierarchical errors was too detailed and specific. On the other hand, they mentioned that most of the features they have judged unnecessary could be interesting for researchers for understanding language patterns in general. They thought this view would be unsuitable for students and learners to use directly.

The **linguists** were all researchers in language acquisition. Most stated they process and analyze large corpora as part of their pipeline, and some of the methods mentioned to perform this analysis were feature extraction, distributional statistics analysis, spreadsheets and heatmaps. Similarly to the language tutors, they seemed initially overwhelmed when presented with H-Matrix for exploration, but after interacting with it, said it feels easy to use.

For the text view, linguists suggested to fade out tagged errors from categories which were filtered-out. Another suggestion was to display a visual representation summarizing the errors in one essay (similar to the details showed on hover, but permanently on screen), this can help to spot interesting essays to analyze. They also were pleased that this view allowed them to validate the feature extraction.

Regarding the hierarchical matrix view, they found it to be useful to compare groups of learners. One participant mentioned: *"It [matrix view] would be interesting to compare naturalistic learners with instructed learners."* On the other hand, they said the category names seem arbitrary and specific, which requires the linguist to be familiar with them in order to fully take advantage of the tool. In addition, most linguists wanted to use their own set of features and documents in H-Matrix. The tasks supported were appropriate and they saw it being helpful in their research and data analysis.

They suggested some interactions for the nodes, such as, filtering the essays by clicking at any level of the sub-trees, and to be able to expand or collapse all the children under a node.

In general, they were excited for adopting the tool in their pipeline to look at their own favorite feature phenomena and data patterns. They were particularly fond of the essay view as a mean to validate their feature extraction. They also mentioned that besides linguists, students can benefit from it by having their language teachers use the tool to explore and understand transfer effects.

6 Discussion

Our initial design challenge was to combine hierarchical categorization of linguistic features with distributional statistical analysis across learner groups. The result was to integrate a tree and a matrix visualization. For on-demand changes on the abstraction level in the feature categories, we allow users to aggregate and re-weight rows dynamically in the matrix without having to reprocess the data. The essay view enables lower level exploration, where users can extract text level examples, as well as, validate the feature extraction itself.

From the expert evaluation, we realized that linguists and language tutors have similar goals to understand and identify transfer effects to help second language learners. However, their methodologies greatly differ. Language tutors are mainly interested in smaller groups of essays. For this reason, we would need to design a third view for a smaller set of essays. Perhaps, create profiles for learners displaying extracted errors history, and similarity to larger learner groups, *e.g.* first language, proficiency. Regarding the granularity of the error categories, we can allow tutors to categorize and extract their own textual features using regular expressions.

In contrast, linguists had a positive response and seemed genuinely interested in using H-Matrix to explore their own corpora. We compiled a list of general functionality improvements from both groups of experts, such as, better labeling and color legends, and from linguists regarding the tree interaction.

To tackle some of the feedback, we made a number of changes in the tool. We initially had the matrix and essay view on the same page. However, we noticed that participants rarely interacted with both at the same time. They preferred to explore each view separately as part of their personal workflow. For this reason, we decided to separate the views. The matrix view can still be used to filter and navigate to the essay view, but they are on two separate tabs, which gives them both more screen space and reduces the visual clutter (as shown in the figures). We hope this change will help with the feeling of being overwhelmed the participants mentioned during the study.

In the matrix view, we improved the textual explanations for the color scale legends, we also created the right click menu options on the tree component for filtering the essay level view and options to expand/collapse all levels of a node to improve feature exploration. On the essay view, we added all the current filter options and the error column as well as its in-cell visualization. The category filter in particular, addresses the complaints of the categorical color encoding not being memorable on the essay tagged errors.

7 Conclusion and Future Work

This paper presented H-Matrix, a visualization tool created to support cross-linguistic features exploration in large learner corpora. H-Matrix is the result of an interactive design approach where we both worked with our expert collaborators and evaluated a prototype with expert participants. The expert evaluation indicated some limitations, such as, the different expectations between language tutors and linguists, as well as some visual and interactive issues addressed in our discussion. Even though tutors, like linguists, are trying to analyze and understand transfer effects, their method focuses on individual or a much smaller group of learners. H-Matrix was designed to deal with large learner corpora and it is over-powered for language tutors. On the other hand, linguists were excited with the functionalities of the tool. We used their feedback to improve the usability of our tool in the presented version (Figs. 1 and 2), as mentioned in the discussion section.

As future work, we plan to evaluate this version of H-Matrix by using linguists own corpora and extracted features. We will then be able to fairly evaluate the usefulness of H-Matrix with real use cases and application. A third view for managing essays of individuals, or smaller groups of learners should be designed to better accommodate tutors needs based on their feedback and our discussion.

Acknowledgments

The authors wish to thank the reviewers, our colleagues, and domain experts. This work was supported in part by NSERC Canada Research Chairs and a grant from SFB-TRR 161.

References

[1] Y. Bestgen, S. Granger, and J. Thewissen. Error patterns and automatic l1 identification. *Approaching Language Transfer through Text Classification*, pp. 127–153, 2012.

[2] R. Blanch, R. Dautriche, and G. Bisson. Dendrogramix: A hybrid tree-matrix visualization technique to support interactive exploration of dendrograms. In *IEEE Pacific Visualization Symp. (PacificVis)*, pp. 31–38. IEEE, 2015.

[3] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. Toefl11: A corpus of non-native English. *ETS Research Report Series*, 2013(2).

[4] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[5] N. F. Fernandez, G. W. Gundersen, A. Rahman, M. L. Grimes, K. Rikova, P. Hornbeck, and A. Maayan. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data*, 4:170151, 2017.

[6] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot. International Corpus of Learner English, 2009.

[7] S. Jarvis and M. Paquot. Native language identification, 2015.

[8] D. A. Keim and D. Oelke. Literature fingerprinting: A new method for visual literary analysis. In *Proc. of the IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pp. 115–122, 2007.

[9] E. Kochmar. Error detection in content word combinations. Technical report, University of Cambridge, Computer Laboratory, 2016.

[10] E. Kochmar and E. Shutova. Modelling semantic acquisition in second language learning. In *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 293–302, 2017.

[11] T. Mayer, C. Rohrdantz, M. Butt, F. Plank, and D. Keim. Visualizing vowel harmony. *Linguistic Issues in Language Technology*, 4(2):1–33, 2010.

[12] C. Nobre, N. Gehlenborg, H. Coon, and A. Lex. Lineage: Visualizing multivariate clinical data in genealogy graphs. *IEEE Trans. Vis. Comput. Graphics*, 25(3):1543–1558, 2018.

[13] C. Nobre, M. Streit, and A. Lex. Juniper: A tree+ table approach to multivariate graph visualization. *IEEE Trans. Vis. Comput. Graphics*, 25(1):544–554, 2019.

[14] T. Odlin. Crosslinguistic influence in second language acquisition. *The Encyclopedia of Applied Linguistics*, 2013.

[15] LanguageTooler GmbH. LanguageTool API - spell and grammar checker. URL: https://languagetool.org/. Accessed in June 2019.

[16] C. Rohrdantz, M. Hund, T. Mayer, B. Wälchli, and D. A. Keim. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. In *Computer Graphics Forum*, vol. 31, pp. 935–944. Wiley Online Library, 2012.

[17] D. Sacha, Y. Asano, C. Rohrdantz, F. Hamborg, D. Keim, B. Braun, and M. Butt. Self organizing maps for the visual analysis of pitch contours. In *20th Nordic Conf. of Computational Linguistics*, pp. 181–189, 2015.