A Bayesian Approach to Inferring Prerequisite Structures and Topic Difficulty in Language Learning

Anh-Duc Vu,[†‡] **Jue Hou,**[†‡] **Anisia Katinskaia,**[†‡] **Ching-Fan Sheu,**[◇] **Roman Yangarber**[‡]
[†]Department of Computer Science, University of Helsinki, Finland
[‡]Department of Digital Humanities, University of Helsinki, Finland
[◇] National Cheng Kung University, Taiwan
first.last@lhelsinki.fi

Abstract

Understanding how linguistic topics are related to each another is essential for designing effective and adaptive second-language (L2) instruction. We present a data-driven framework to model topic dependencies and their difficulty within a L2 learning curriculum. First, we estimate topic difficulty and student ability using a three-parameter Item Response Theory (IRT) model. Second, we construct topic-level knowledge graphs—as directed acyclic graphs (DAGs)—to capture the prerequisite relations among the topics, comparing a threshold-based method with the statistical Grow-Shrink Markov Blanket algorithm. Third, we evaluate the alignment between IRT-inferred topic difficulty and the structure of the graphs using edge-level and global ordering metrics. Finally, we compare the IRT-based estimates of learner ability with assessments of the learners provided by teachers to validate the model's effectiveness in capturing learner proficiency. Our results show a promising agreement between the inferred graphs, IRT estimates, and human teachers' assessments, highlighting the framework's potential to support personalized learning and adaptive curriculum design in intelligent tutoring systems.

1 Introduction

A key goal of Intelligent Tutoring Systems (ITS) is to support *personalized* learning by answering key questions: What does the student know? How are they performing? What should they learn next? Achieving this requires three components: a Domain Model (to represent subject knowledge), the Student Model (to represent learner proficiency), and the Instruction Model (to implement pedagogical strategy). Of these, the domain model is foundational, as it informs both the student assessment and the instructional choices. Prior work has explored domain modeling in many learning domains, such as mathematics (Ritter et al., 2007; Arroyo et al., 2014; Klinkenberg et al., 2011).

Beyond proficiency estimation, recent work emphasizes domain models that offer pedagogical insights—such as relative topic difficulty and efficient or optimal learning paths (Swamy et al., 2022; Cohausz, 2022; Weidlich et al., 2022). These can help teachers adapt instruction and improve learning outcomes. In this paper, we focus on modeling relationships among *topics* in language learning using two approaches: predictive modeling and causal modeling. The causal model aims to provide an interpretable domain structure, while the predictive model offers empirical estimates of learning outcomes.

We collect data from real-world learners in our language learning system, Revita (Katinskaia et al., 2018; Katinskaia and Yangarber, 2018; Katinskaia et al., 2017).[1] In Revita's learning setting, learners complete exercises related to grammar topics in the target language. These exercises are automatically generated from texts that learners upload themselves or select from a shared library of materials. The exercises are presented in the form of multiple-choice or fill-in-the-blank ("cloze") questions. Each question is associated with one or more *learning topics*—a.k.a. linguistic constructs (Katinskaia et al., 2023)—and the learner's answer is graded according to its correctness in terms of each topic. We collaborate with language teachers from several universities and collect real data from language learners.

The main goal of this paper is to explore the domain model—using data from learners of Russian, one of several languages offered by the Revita learning platform—which is based on the Russian topics and their relationships. We highlight the following contributions of this paper:

1. We present a simple causal modeling scheme for the domain model and model topics with a directed acyclic graph (DAG). The nodes

[1] revitaai.github.io

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 737–751
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

in the graph represent topics, and the edges represent the relationships between them.

2. We verify our graph structure with predictive analysis: Bayesian network and hierarchical item response theory (IRT) model.

The paper is organized as follows. In section 2 we outline relevant prior work. Section 3 describes our topic inventory, the process of data collection and performance aggregation by topic. Section 4 describes our approach to build the prerequisite graph structures and the statistical models we use to verify the graph structures. Section 5 shows the experiment results. Section 6 concludes with current directions of research.

2 Related Work

Several approaches for modeling learning have been proposed. We briefly review two types of models: (1) predictive models and (2) causal models.

Predictive models focus on predicting with a set of independent latent variables. When modeling learning, these latent variables refer to the levels of the student's proficiency on various learning topics. One approach is Item Response Theory (van der Linden and Hambleton, 2013). ITS is not the only application of IRT—it can be applied in many settings, including stress testing, psychological and medical testing, etc. Depending on the application domain, the latent trait can be level of anxiety, neurosis, authoritarian personality, etc. IRT has an information-theoretic basis similar to "Elo" ratings (Elo, 1978). The Elo formulas, originally developed for rating chess players, have been adapted in the context of ITS (Pelánek, 2016; Hou et al., 2019). The language-learning domain is more complex than other domains where IRT is used, since the learning topics to be mastered are relatively much more numerous, and have complex relationships among them.

With the rise of deep learning in recent years, *deep knowledge tracing* (DKT) was proposed (Piech et al., 2015), modeling the state of learner knowledge with a recurrent neural network—RNN (Hochreiter and Schmidhuber, 1997). Researchers have proposed several neural network-based approaches (Zhang et al., 2017; Abdelrahman and Wang, 2019; Su et al., 2018; Liu et al., 2019; Pandey and Srivastava, 2020; Song et al., 2021). The benefit of applying neural networks is that they do not require human-engineered features; despite the success of deep learning, they suffer from a lack of interpretability (Jiang et al., 2024).

Causal models describe the causal relationships in a system. In our case, we consider the causal relationship to be the *prerequisite* relationship among topics or the learner's knowledge states. The benefit of using causal models is that they can provide a more directly interpretable representation of the domain knowledge (Jiang et al., 2024). Some causal models describe the domain as a directed acyclic graph (DAG), which provides direct value from the perspective of pedagogy. Researchers have explored the use of causal models in education with Bayesian networks (Pardos and Heffernan, 2010) or Markov Blanket (Jiang et al., 2024).

In the field of education, Knowledge Space Theory (KST) (Doignon and Falmagne, 2012) can also be considered as a special graphical causal model. KST is a mathematical framework for modeling the learner's knowledge, and represents the learner's current proficiency as a set of mastered skills, which is referred as a *knowledge state*. Each state contains a subset of the skills in the domain. The student has mastered the domain when she reaches the state containing all skills. KST models not only the learner knowledge, but also learning paths, starting from the empty set toward the full set of topics. Various approaches are used to build a knowledge space, from explicit elicitation of knowledge from human experts to data-driven methods, such as Formal Concept Analysis (FCA) (Ganter and Wille, 2012).

3 Data

This work uses learner data collected in collaboration with language teachers at several universities. The dataset covers university students learning Russian as a second language (L2), whose levels range from A1 to C2 on the CEFR scale (Little, 2007), both as part of their university courses and as independent study. Learners upload texts of personal interest, or, if participating in a university course, practice with texts selected or adapted for them by their teachers. Based on the selected texts, the Revita intelligent language tutoring system automatically generates interactive exercises (Katinskaia et al., 2023).[2]

Revita supports a variety of exercise types for each language, including grammar, vocabulary, lis-

[2] revita.helsinki.fi

Topics	Examples
(1) Verb: II conjugation	Мы скоро увидим восход. (We **will see** the sunrise soon.)
(2) Complex pronoun:	Нам нужно кое о чем поговорить. (We need to talk **about something**)
(3) Perfective vs. imperfective aspect	Страны согласовали проект о будущих отношениях. (The countries **agreed on** a draft on future relations.)
(4) Dative subject with predicative adjective, or with impersonal verb	Мне необходимо поговорить с врачом. (**I** need to talk to a doctor. Literally: [it is] necessary **for me** to talk to a doctor.)

Table 1: Examples of instances of *topics* found in text (underlined). ***Candidates*** are words that will be chosen for exercises about the topics (marked in **bold**).

tening comprehension, etc. It also provides continual diagnostic assessment. It assists the learners with contextualized feedback and hints depending on their answers. Exercise creation and hint generation are built upon a linguistically-informed domain model, which drives the personalized selection and generation of exercises based on each learner's proficiency level. In this study, we focus on learner data from grammar exercises in Russian, which serve as the foundation for modeling topic dependencies and estimating topic difficulty.[3]

3.1 Data collection

Topics: In this paper, we use the term *topic* to refer to specific language learning targets (also known as "skills" in ITS and education literature)—for example, particular patterns of nominal case usage, verb conjugation classes, syntactic constructions involving negation and tense, etc. These are not simply individual grammatical features, such as *past tense* or *plural number*, but rather combinations that reflect in a meaningful fashion how language is taught and learned. For instance, learners may work on mastering topics such as *past tense of a certain verbal paradigm*, rather than *past tense* in general.

To define these topics, we consulted with experts in language pedagogy and textbooks, to align with real-life instructional goals. Table 1 shows examples of topics and exercises that target them.

Exercises: All exercises are automatically generated by the Revita system, based on authentic texts chosen by the teachers and learners from arbitrary sources. The system creates a number of exercise types; here we focus on fill-in-the-blank ("cloze") and multiple-choice exercises. In a cloze exercise, the system hides certain words or phrases, and shows the learner a hint—the lemma (dictionary form) of the hidden word or phrase. The learner's task is to enter the correct surface forms, based on the context of the cloze. In a multiple-choice exercise, the learner is given several options to choose from, with the options generated automatically.

Learners are allowed multiple attempts for each exercise. When an answer is incorrect, the system provides hints on subsequent attempts to support the learner. These hints *gradually* guide the learner toward the correct answer—starting with general guidance and becoming increasingly specific with each additional attempt.

The exercise sequencing strategy follows a hybrid adaptive design. The system is designed to model the learner's state to select those exercises that optimally match each learner's current proficiency—targeting an expected success rate of 50%, to keep the exercises appropriately challenging. This is in keeping with Vygotsky's theory of the Zone of Proximal Development, which states that for optimal learning, the exercises must not be too difficult too often (to avoid frustrating the learner) and not too easy too often (to avoid boring the learner) (Poehner, 2008). Alternatively, learners can manually select their own study paths using a predefined lesson structure, organized from easier to more difficult topics.

Assignment of credit and penalty: Each exercise is associated with one or more topics. The system evaluates the learner's response to estimate performance on each topic individually. A response may be correct with respect to some topics but incorrect with respect to others—for instance, a learner might use the correct verb tense but the wrong grammatical person. If the learner answers correctly only after receiving hints, we apply a slight penalty, proportionally distributed across the topics linked to those hints. To assign credit and penalty, the system uses several NLP components, including a morphological analyzer, dependency parser, and rule-based pattern matcher. These tools compare the learner's response with the correct an-

[3] All learner data was anonymized prior to analysis in accordance with ethical research requirements and standards.

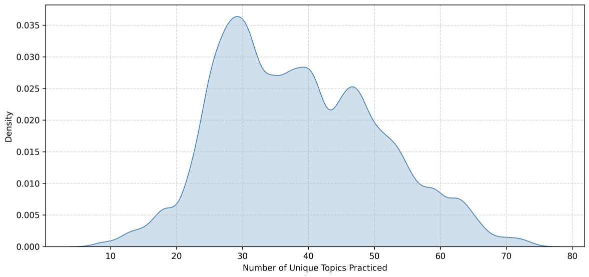

Figure 1: Distribution of the number of unique topics practiced by each student.

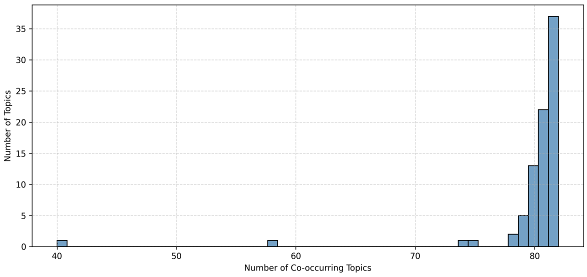

Figure 2: Distribution of counts of co-practiced topics, based on shared student activity. Each value on the X-axis indicates with how many other topics each topic was co-practiced. The Y-axis shows how many topics have the given co-practiced count.

swer to determine the topic-level performance for each exercise.

Assignment of credits and penalties is one of the main challenges in our work on assessment. Most statistical approahes, such as IRT, have a clear definition of an *item*, and a clear credit standard—right/wrong answer given by the learner in response to the item. The classic example of an item in IRT is a test question, e.g., in mathematics: it is dichotomous and unambiguous, with a clear judgment of the answer—correct or incorrect. Our major challenge is that our topics are not judged directly, as test items are in other learning domains. It is challenging to determine the credit and penalty for each topic based on the learner's answer, because the link from exercise to topic is *one-to-many*. This one-to-many nature of the link makes the standard of credit less clear. To tackle this problem, a more sophisticated approach is required to assign credit and penalty. We also face another common problem in language learning and assessment: ambiguity. A substantial proportion of exercises admit *more than one* possibly correct answer, leading to the problem of determining grammatical correctness (Katinskaia and Yangarber, 2021, 2023, 2024). The quality of our NLP components directly impacts the accuracy of the assessment, and therefore the quality of our learning data.

3.2 Data pre-processing

We have collected over 470K student exercise attempts, each with credit and penalty assigned. These exercises were completed by 1,639 unique students. These exercises span over 200 detailed grammatical constructs (Katinskaia et al., 2023), which we group into a smaller set of learning topics that align with pedagogical learning targets, as described above. From this, we derive over 80 distinct topics to be used for modeling and construction of prerequisite graphs.

Each exercise is associated with one or more linguistic topics. To enable topic-level analysis, we "explode" (i.e., multiply out) the data, so that each exercise attempt is represented *multiple* times—once per each topic linked to the exercise. This allows us to track student performance separately for each topic. The number of "exploded" data points—pair-wise records linking between student and topic—is approximately 990K.

The histogram in Figure 1 shows the distribution of unique topics practiced per student. Most students engage with 25 to 50 distinct topics, with a concentration around 30. Since learners tend to focus on topics appropriate to their proficiency level, we expect considerable overlap in practiced topics among students of similar levels. This local overlap is useful for constructing prerequisite graphs, as it provides aligned performance patterns across comparable learners without requiring complete topic coverage by each individual.

We next check what topics are *co-practiced* with other topics—i.e., which topics have been practiced together with other topics by at least one student. Figure 2 shows how many topics are co-practiced with other topics. In fact, most topics are co-practiced with 80 or more other topics, indicating a highly interconnected curriculum, where students tend to practice multiple topic combinations. This highlights the dense overlap in student exposure across topics, which is a useful signal for data-driven construction of dependency graphs.

Figure 3 shows the distribution of students that have engaged with each topic. While some topics are widely practiced by hundreds of learners, others are encountered by only a few students, indicating potential variation in topic popularity, curriculum coverage, or personalized learning paths. This vari-

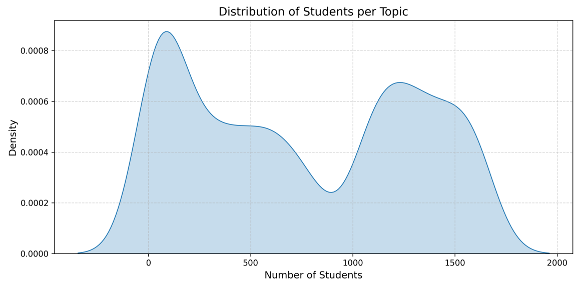

Figure 3: Distribution of the number of students per topic.

ability may impact both topic-level estimation and the structure of the prerequisite graph.

We have also collected data from over 50 students who completed 100 or more exercises each, and who have teacher-assigned CEFR levels. This subset provides a valuable reference for evaluating student ability and topic difficulty. Overall, the dataset's size and structure support both robust probabilistic modeling of learner proficiency and detailed analysis of topic interdependencies.

4 Methodology

4.1 Prerequisite graph construction

To represent the prerequisite structure of the topics, we construct a DAG (Chickering, 2002) over learning topics, where each node represents a topic. Directed edges in this graph indicate prerequisite relationships inferred from empirical student performance patterns. Specifically, a directed edge from topic A to topic B, denoted $A \rightarrow B$, means that mastery of topic A is likely a prerequisite for success on topic B.

We explore multiple methods for constructing the prerequisite graph. The first is a threshold-based approach, in which a directed edge $A \rightarrow B$ is added if a statistically significant fraction of students consistently perform better on topic A than *the same students* perform on topic B. This approach focuses solely on relative performance outcomes across topics. By aggregating student-specific accuracy rates, the method infers likely learning dependencies under the assumption that prerequisite topics are easier for students to master than their dependents.

The second method is Grow-Shrink Markov Blanket (GS-MB) approach to learn topic dependencies based on statistical conditional independence tests (Margaritis and Thrun, 2000). We first identify potential neighbors of a target topic by evaluating unconditional correlations (grow phase), then we remove far neighbors by testing for conditional independence given the remaining set (shrink phase) until reaching the actual Markov blanket of the topic. The resulting undirected dependencies are then converted into directed edges using edge orientation heuristics.

To ensure that the resulting prerequisite graph is a valid directed acyclic graph (DAG), we apply data-driven postprocessing to eliminate cycles and resolve bidirectional edges. If a cycle is detected, we iteratively remove the weakest edge within the cycle—where "weakness" is determined using statistical evidence such as a low agreement ratio or minimal co-occurrence frequency across student performance data. Unlike the traditional Grow-Shrink approach proposed by Margaritis and Thrun (2000), which attempts to reverse and reinsert removed edges followed by directional propagation heuristics, our method permanently removes low-confidence edges without reorientation. This simplification focuses on preserving only the most statistically supported links while enforcing global acyclicity. For bidirectional dependencies (i.e., both $A \rightarrow B$ and $B \rightarrow A$), we retain only the edge with the stronger statistical support, ensuring a consistent and interpretable prerequisite structure.

4.2 IRT Modeling of Student Performance

We use a probabilistic model to estimate student ability and topic difficulty based on their exercise-performance data. Specifically, we apply the three-parameter logistic (3PL) Item Response Theory (IRT) model (Baker, 2001). In 3PL, each student has an ability parameter θ_s, and each topic has two parameters: difficulty β_t, and discrimination α_t. We also take into account the factor of luck as guessing parameter g. The probability that a student s answers topic t correctly is modeled as:

$$c_{u,t} \sim \text{Bernoulli}\left(g + (1-g) \cdot \sigma\left(\alpha_t(\theta_s - \beta_t)\right)\right)$$

where $\sigma(\cdot)$ is the sigmoid function.

We assume a fixed guessing parameter $g = 0.01$ for cloze-style exercises, which approximates the probability of answering correctly by chance. For multiple-choice exercises, g is determined dynamically based on the number of answer options.

To estimate the posterior distributions of the model parameters, we perform fully Bayesian inference via Markov Chain Monte Carlo (MCMC) (Gilks et al., 1995), using the No-U-Turn Sampler

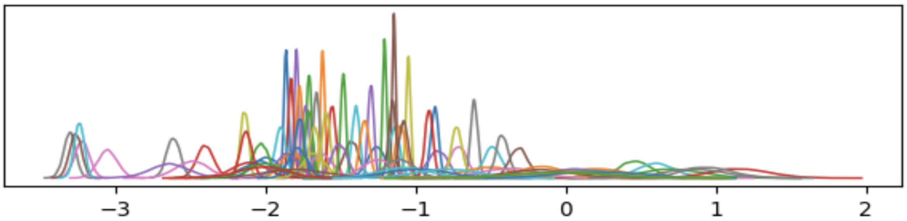

Figure 4: Posterior distributions of topic difficulty (β) estimated by the IRT 3PL model. X-axis is estimated topic difficulty. Each curve represents the density for a different topic.

(NUTS) (Hoffman and Gelman, 2014) as implemented in PyMC (Patil et al., 2010; Salvatier et al., 2016).

NUTS is a gradient-based sampling algorithm that extends Hamiltonian Monte Carlo (HMC) by adaptively deciding how many steps to simulate on each iteration, based on the gradients of the log-posterior. It dynamically simulates forward and backward trajectories in the parameter space, stopping when a "U-turn" is detected, and then selects a new sample from the visited states using a probability distribution. The posterior samples of β and α for each topic capture both the parameter estimates (mean) and their associated *uncertainty* (standard deviation), enabling more detailed downstream analysis and validation of the graph structure.

4.3 Comparing Graph Structure with IRT Difficulty

We assess the extent to which the structure of the prerequisite graph agrees with IRT-inferred topic difficulty. Intuitively, if topic A is a prerequisite for topic B, then A should be easier (i.e., have lower β) than B. To evaluate this alignment, we use three complementary metrics.

Edge Agreement Score (EAS) measures the proportion of edges in the graph that follow the expected difficulty order. For each edge $A \to B$, we check whether $\beta_A < \beta_B$. The EAS is calculated as the fraction of such edges over all edges in the graph. A perfect score of 1.0 indicates that *all* edges point from an easier to a harder topic.

Weighted Direction Score (WDS) refines this idea by incorporating the size of the difficulty gap. Rather than using a hard threshold, we score each edge $A \to B$ using a sigmoid-transformed difference between the difficulties of topics A and B:

$$\sigma(\beta_A, \beta_B) = 1/(1 + e^{-(\beta_B - \beta_A)})$$

This yields higher scores when β_B is much greater than β_A, and values near 0.5 when the difference is small or uncertain. WDS offers a smoother estimate that rewards clear hierarchical structure.

Kendall's Tau, originally introduced by Kendall (1938), is designed to measure the ordinal association between two ranked variables. We use it to compare the global ordering of topics implied by the graph with the ranking induced by the IRT-inferred difficulty estimates. This is done by computing a topological sort of the graph to obtain a linear topic ordering, which is then correlated with IRT's β values using Kendall's Tau. A high Tau value indicates strong agreement: topics that appear earlier in the graph tend to be easier than those ranked later.

Together, these metrics offer both local and global perspectives on how well the learned DAG structure matches the IRT-inferred difficulty landscape.

5 Experiments and Results

5.1 IRT Estimations

Figure 4 shows the posterior density of all topic difficulty estimates. The IRT model estimates topic difficulty values ranging from -3.31 to 1.16, giving a total range of 4.47 on the X-axis. Of 83 topics, 38 have standard deviations below 0.05, and 55 are below 0.10, meaning that their difficulty estimates are quite stable. For 95% confidence intervals (CI), 50 topics (∼60% of all topics) have interval widths under 0.30, which is about 6.7% of the full difficulty range. For 17 topics (20% of all topics), the estimated CI width is under 0.10—only 2.2% of the full range. These statistics suggest that many topics are estimated with high confidence, and they are reliable enough to be used for comparison with the topic graph.

Figure 5 illustrates the relationship between uncertainty in topic difficulty (standard deviation of β) and topic discrimination (mean of α). Topics with higher discrimination α tend to show lower

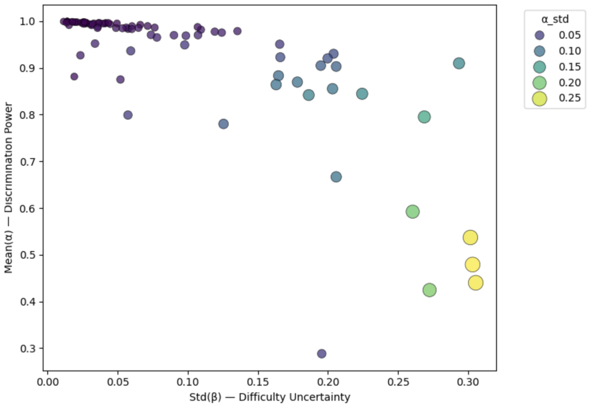

Figure 5: Correlation between uncertainty in the difficulty of a topic—std(β)—and discrimination parameter of the topic—mean α—for each topic. Each point represents a topic. Marker color and size indicate the *uncertainty* in α.

uncertainty in their estimated difficulty, indicating more stable and informative estimates. In contrast, topics with lower discrimination show greater uncertainty in β, as well as higher variability in α, as indicated by larger, lighter-colored markers. This pattern suggests that strongly discriminative topics provide more reliable signals for modeling.

A detailed heatmap of the correctness of the student responses by topic and student ability quantiles (Q1 = lowest, Q5 = highest) is shown in Appendix Figure 9 and 10. In the heatmap, topics are ordered by their IRT-estimated mean difficulty β. The color of each cell shows the average correctness rate, and the number of student-topic interactions. As expected, higher-ability students (Q4–Q5) perform better, particularly on the more difficult topics, reinforcing the validity of the estimated difficulty scores.

Figure 6 shows how the estimates of student ability θ vary across CEFR levels assigned by the teachers. The correlation between CEFR grade and IRT-estimated ability is moderate, with a Spearman coefficient of $r = 0.473$, indicating that as CEFR level increases, IRT-based ability estimates also tend to rise.

Figure 7 shows the relationship between the number of exercises completed by each student and the uncertainty in their estimated ability, measured as the posterior standard deviation of θ. There is a strong negative correlation ($r = -0.758$, $p < 0.001$), indicating that students who complete more exercises tend to have more confident (lower-variance) ability estimates. This supports the intuitive notion that additional observations reduce posterior uncertainty in the IRT model.

Figure 8 shows a strong negative relationship between the number of students who practiced a topic and the uncertainty in that topic's IRT difficulty estimate. Topics attempted by more students tend to have significantly lower standard deviation in their β values, suggesting higher confidence in the estimated difficulty. This trend is quantitatively supported by a Spearman correlation of $r = -0.899$ ($p < 0.001$), confirming that broader student coverage leads to more stable parameter estimates.

5.2 Graph construction

We construct two types of topic prerequisite graphs to capture learning dependencies. The first, a threshold-based graph, connects topics where a consistent performance advantage suggests one precedes the other. The second, built using the Grow-Shrink Markov Blanket algorithm, identifies conditional dependencies between topics based on statistical independence tests.

The threshold-based graph includes 83 nodes and 173 edges, resulting in a wide and dense structure with many inferred prerequisite links. In contrast, the GS-MB graph is sparser, with 80 nodes and 86 edges, forming a deeper and narrower hierarchy. Both graphs are processed to remove cycles and bidirectional edges, ensuring they are valid directed acyclic graphs (DAGs). Visualizations of both graphs can be found in the Appendix (Figures 11 and 12).[4]

Both graphs offer useful perspectives. When we manually examine their qualitative plausibility from a linguistic standpoint, we find that the threshold-based graph often aligns more intuitively with expected topic relationships in Russian, suggesting that threshold-based edges may capture pedagogically meaningful dependencies more effectively than the GS-MB structure. We will explore this in further depth in future work.

5.3 Graph vs. IRT estimations

Two approaches are evaluated for constructing topic prerequisite graphs: a threshold-based method and the Grow-Shrink Markov Blanket algorithm. Both produce DAGs, which are evaluated for alignment with the IRT-inferred topic difficulties using three metrics: Edge Agreement Score (EAS),

[4] Both of these graphs are too large to fit into the paper; please see the complete graph of threshold-based approach here and GS-MS approach here.

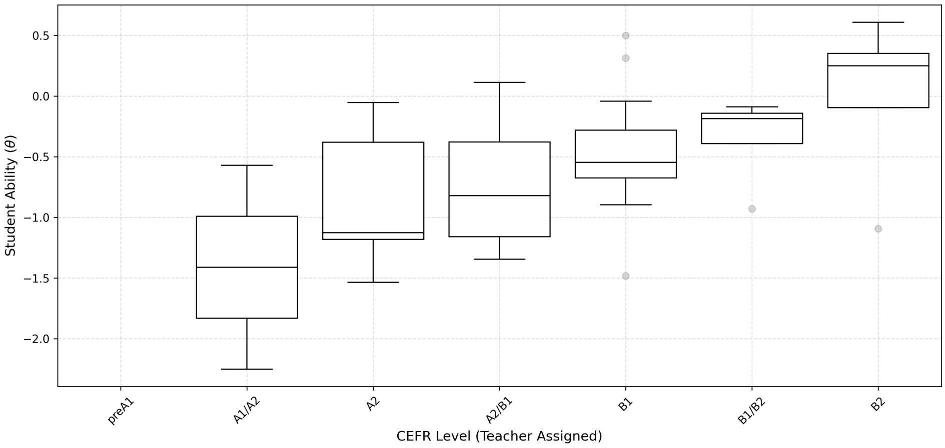

Figure 6: IRT ability estimates θ across teacher-assigned CEFR levels. Boxplot shows distribution of student abilities per CEFR level.

Graph	**Uncertainty Filter**	**EAS**	**WDS**	**Kendall's Tau**
Threshold-based (G_{dag1})	$\sigma_\beta < 0.05$	**1.000**	**0.869**	**0.240**
	$\sigma_\beta < 0.10$	0.604	0.609	0.220
GS-Markov Blanket (G_{dag2})	$\sigma_\beta < 0.05$	0.523	0.514	0.050
	$\sigma_\beta < 0.10$	0.505	0.502	0.014

Table 2: Agreement between graph structure and IRT-estimated topic difficulty. EAS: Edge Agreement Score. WDS: Weighted Direction Score. Kendall's Tau compares topological sort with IRT difficulty rank.

Weighted Direction Score (WDS), and Kendall's Tau.

Table 2 summarizes the results under two topic uncertainty thresholds—$\sigma_\beta < 0.05$ and $\sigma_\beta < 0.1$—which correspond to subsets of 30 and 50 topics, respectively. The threshold-based graph consistently shows stronger alignment with IRT difficulty estimates, achieving perfect edge agreement (EAS = 1.00), high directional consistency (WDS = 0.869), and a moderate Kendall's Tau of 0.240 under stricter filtering. Even with relaxed thresholds, it maintains relatively strong scores across all three metrics. In contrast, the GS-MB graph produces lower EAS, WDS, and notably near-zero Kendall's Tau values (i.e., 0.050 and 0.014), indicating that its topological structure does not match the global difficulty ranking well.

These results suggest that while GS-MB could be effective at capturing local conditional dependencies, it falls short in representing an overall difficulty hierarchy—a strength more consistently captured by the threshold-based method.

6 Conclusion

In this work, we present a unified framework for modeling topic difficulty and learning dependencies in second-language acquisition, leveraging large real-world learner data from thousands of students. Using probabilistic modeling and graph-based structure learning, we analyze over 470K student exercise attempts spanning more than 80 topics. Our aim is twofold: (1) to estimate topic-level difficulty and learner ability using a Bayesian IRT model, and (2) to construct interpretable prerequisite graphs that reveal topic hierarchies potentially useful for improving learning.

We compare two graph construction methods: a threshold-based approach that aggregates relative performance gaps across students, and a Grow-Shrink Markov Blanket (GS-MB) method based on statistical conditional independence tests. Three evaluations using Edge Agreement Score (EAS), Weighted Direction Score (WDS), and Kendall's Tau show that the threshold-based method aligns more closely with the IRT-inferred topic difficulties. This supports the hypothesis that prerequisite topics tend to be easier than their dependents, and suggests that simple, data-driven heuristics can reveal meaningful pedagogical structures.

Our findings also demonstrate that model confidence is strongly influenced by the *volume* and *diversity* of learner data. Students who have completed more exercises tend to have lower uncertainty in their ability estimates; topics practiced by

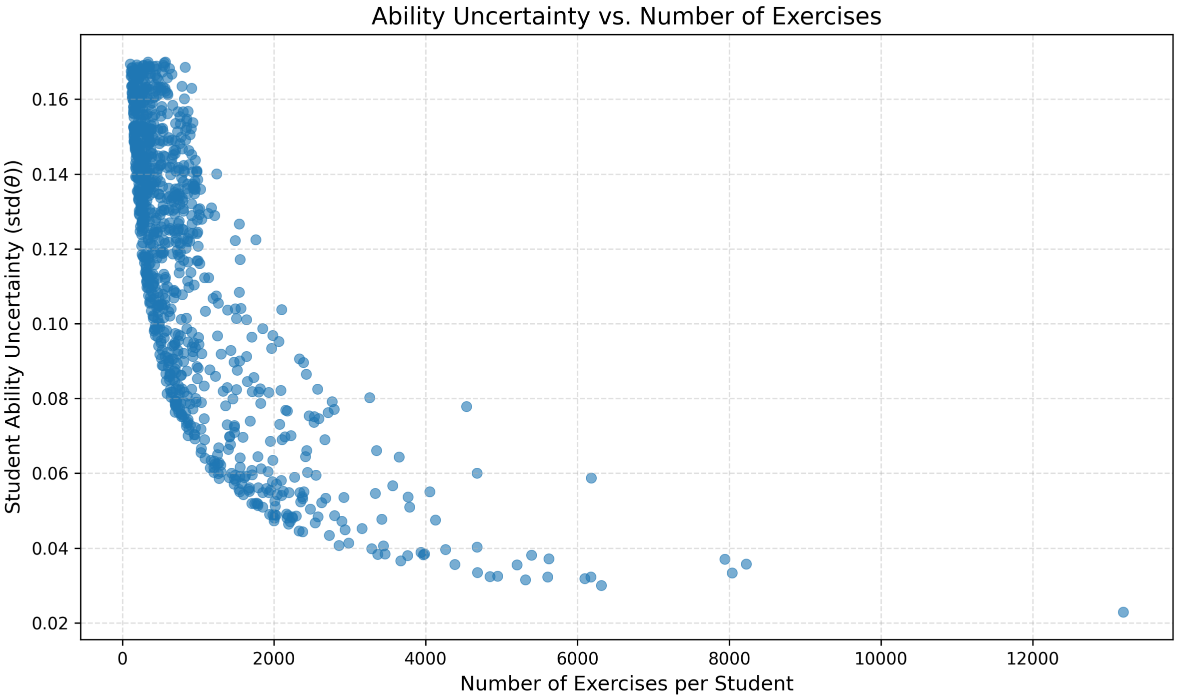

Figure 7: Relationship between number of exercises completed and ability uncertainty (standard deviation of θ). Each point represents a student.

more learners show lower variability in their difficulty estimates. These patterns highlight the value of large-scale learner data in stabilizing the parameter estimates and guiding curriculum analysis.

Moreover, the estimated IRT abilities exhibit a correlation with teacher-assigned CEFR levels, providing external validation for the model and supporting its use in real-world learner assessment. We further explore several aggregate statistics, including topic-topic co-occurrence and student-topic interaction distributions, to explore coverage patterns and the implications for curriculum design.

In summary, this study contributes a robust methodology for combining statistical modeling and graph structure learning in an educational setting. The approach offers practical tools for curriculum designers and language educators to identify learning gaps, and to evaluate learner proficiency. In future work, we will explore extending the model to dynamic learning sequences, fine-grained topic representations, and multilingual adaptation, to further enhance intelligent language tutoring systems.

Limitations

Our results at present have several limitations that may affect the generalizability and precision of the results.

The dataset primarily consists of learners at the A2, B1, and B2 levels, with relatively few samples from C-level students and very limited representation of pre-A1 and A1 learners. As a result, the inferred difficulty hierarchy and student ability estimates may not fully reflect the learning needs or patterns of beginners and advanced learners.

The distribution of labeled performance data is imbalanced: 78.4% of responses are correct, while only 21.6% are incorrect. This skew may reduce the model's sensitivity to detecting subtle topic-level challenges, and can introduce bias in estimating both the topic difficulty and discrimination parameters.

Addressing these gaps—through more diverse learner sampling and more balanced task evaluation—would improve the robustness of future modeling efforts.

References

Ghodai Abdelrahman and Qing Wang. 2019. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 175–184.

Ivon Arroyo, Beverly Park Woolf, Winslow Burelson,

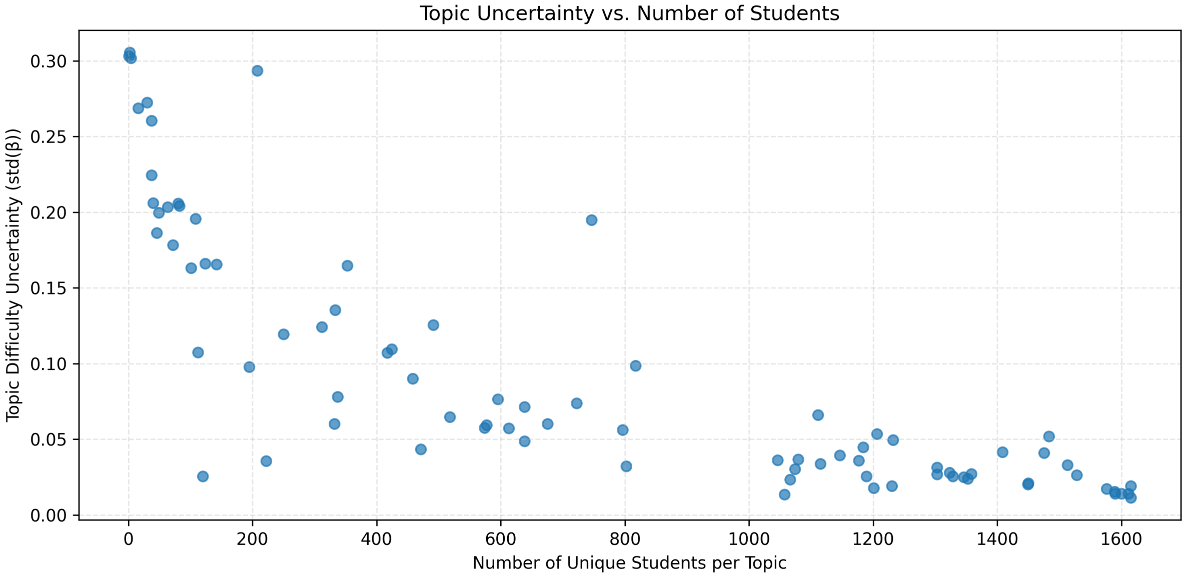

Figure 8: Topic difficulty uncertainty (std(β)) vs. number of unique students per topic. Each point represents a topic.

Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4):387–426.

Frank B. Baker. 2001. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Lea Cohausz. 2022. Towards real interpretability of student success prediction combining methods of xai and social science. *International Educational Data Mining Society*.

Jean-Paul Doignon and Jean-Claude Falmagne. 2012. *Knowledge spaces*. Springer Science & Business Media, New York, NY.

Arpad E. Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub., New York.

Bernhard Ganter and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, New York, NY.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew D Hoffman and Andrew Gelman. 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. In *Journal of Machine Learning Research*, volume 15, pages 1593–1623.

Jue Hou, Maximilian W Koppatz, José Marıa Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, and Roman Yangarber. 2019. Modeling language learning using specialized Elo ratings. In *BEA: 14th Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of Association for Computational Linguistics*.

Bo Jiang, Yuang Wei, Ting Zhang, and Wei Zhang. 2024. Improving the performance and explainability of knowledge tracing via markov blanket. *Information Processing and Management*, 61(3):103620.

Anisia Katinskaia, Jue Hou, Anh-Duc Vu, and Roman Yangarber. 2023. Linguistic constructs represent the domain model in intelligent language tutoring. In *EACL: 17th Conference of European Chapter of Association for Computational Linguistics*.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. In *6th Workshop on NLP for CALL and 2nd Workshop on NLP for Research on Language Acquisition, at NoDaLiDa*, Gothenburg, Sweden.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Anisia Katinskaia and Roman Yangarber. 2018. Digital cultural heritage and revitalization of endangered Finno-Ugric languages. In *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, Helsinki, Finland.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing grammatical correctness in language learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–146.

Anisia Katinskaia and Roman Yangarber. 2023. Grammatical error correction for sentence-level assessment in language learning. In *18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 488–502.

Anisia Katinskaia and Roman Yangarber. 2024. Gpt-3.5 for grammatical error correction. In *Proceedings of COLING-LREC: Joint International Conference on Computational Linguistics and Language Resources and Evaluation*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. 2011. Computer adaptive practice of maths ability using a new item response model for on-the-fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824.

David Little. 2007. The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4):645–655.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115.

Dimitris Margaritis and Sebastian Thrun. 2000. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 12.

Shalini Pandey and Jaideep Srivastava. 2020. Rkt: Relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1205–1214.

Zachary A. Pardos and Neil T. Heffernan. 2010. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, pages 255–266, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anand Patil, David Huard, and Christopher J Fonnesbeck. 2010. Pymc: Bayesian stochastic modelling in python. *Journal of Statistical Software*, 35(4):1–81.

Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Matthew E Poehner. 2008. *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*, volume 9. Springer Science & Business Media, New York, NY.

Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. 2016. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Xiangyu Song, Jianxin Li, Yifu Tang, Taige Zhao, Yunliang Chen, and Ziyu Guan. 2021. Jkt: A joint graph convolutional network based deep knowledge tracing. *Information Sciences*, 580:510–523.

Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Vinitra Swamy, Bahar Radmehr, Natasa Krco, Mirko Marras, and Tanja Käser. 2022. Evaluating the explainers: Black-box explainable machine learning for student success prediction in moocs. In *EDM*.

Wim J van der Linden and Ronald K Hambleton. 2013. *Handbook of modern item response theory*. Springer Science & Business Media, New York, NY.

Joshua Weidlich, Dragan Gašević, and Hendrik Drachsler. 2022. Causal inference and bias in learning analytics: A primer on pitfalls using directed acyclic graphs. *Journal of Learning Analytics*, 9(3):183–199.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*, pages 765–774.

A Appendix

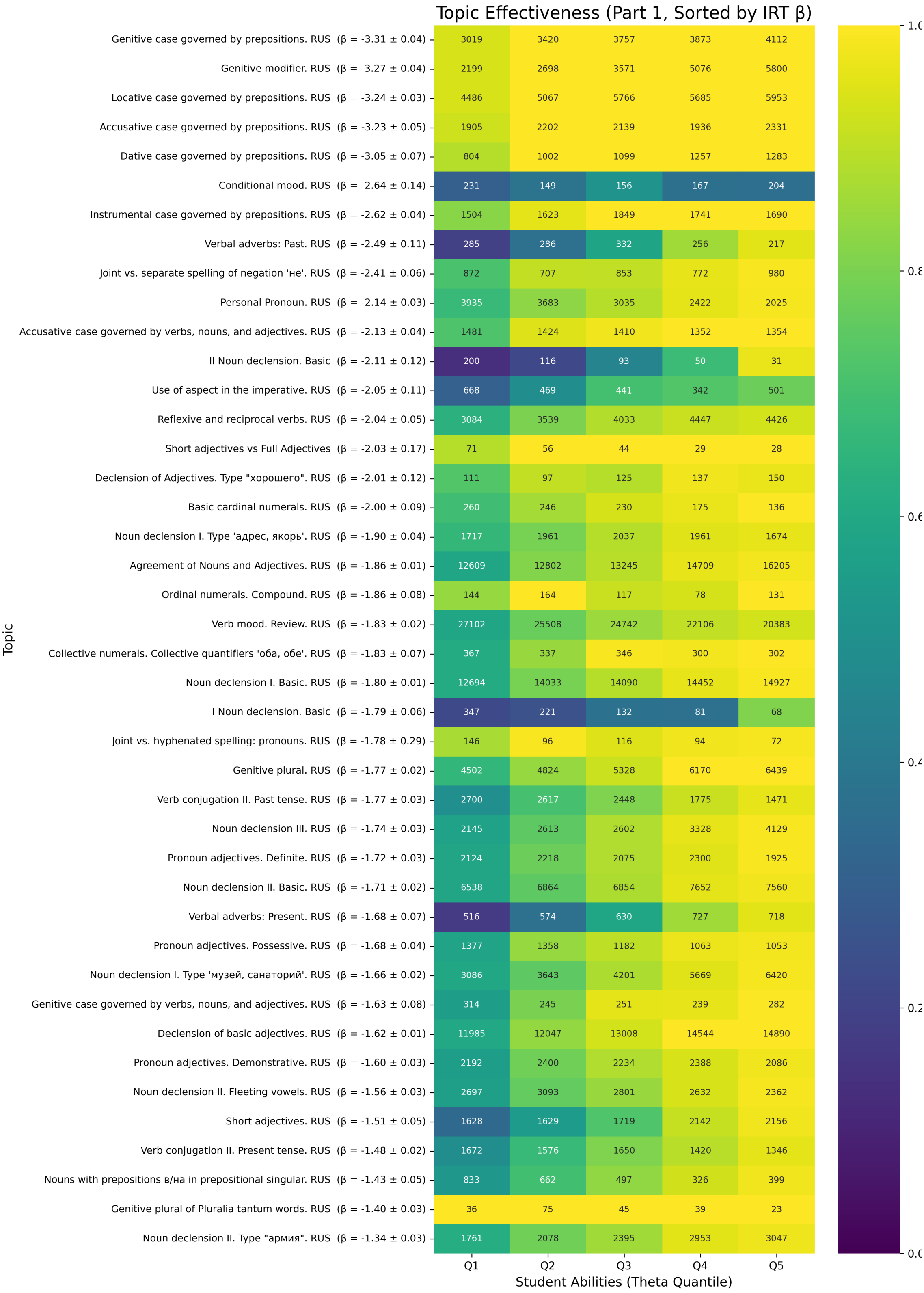

Figure 9: Heatmap of *rate of correct answers* per topic (Part 1). Correctness rates shown per topic and student ability quantile (Q1 = lowest ability, Q5 = highest). Color shows average correctness rate. Number in box indicates support: the number of student-topic interactions. Topics are ordered by their IRT-estimated difficulty β.

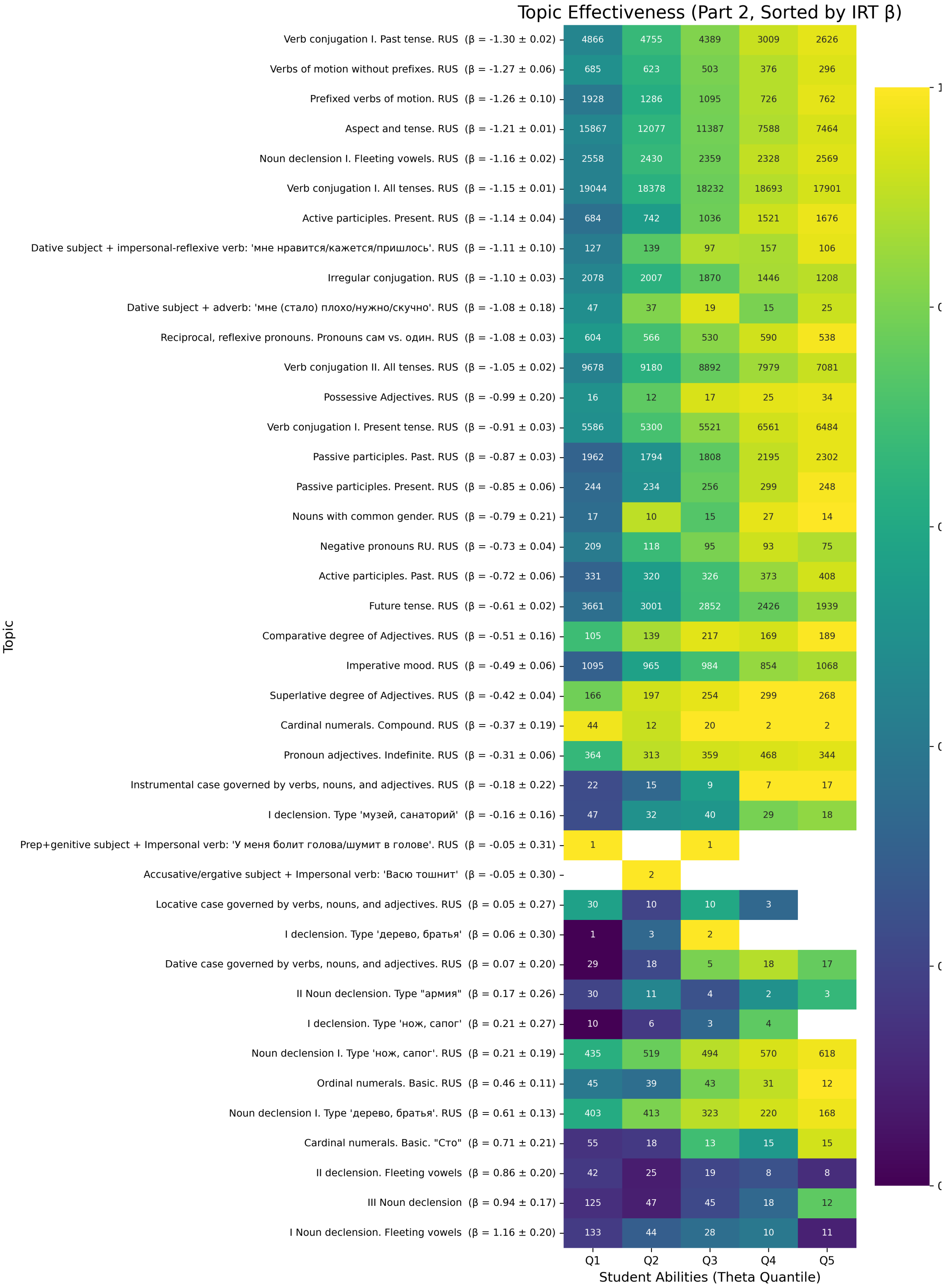

Figure 10: Heatmap of correctness per topic (Part 2); continuation of the heatmap showing remaining topics.

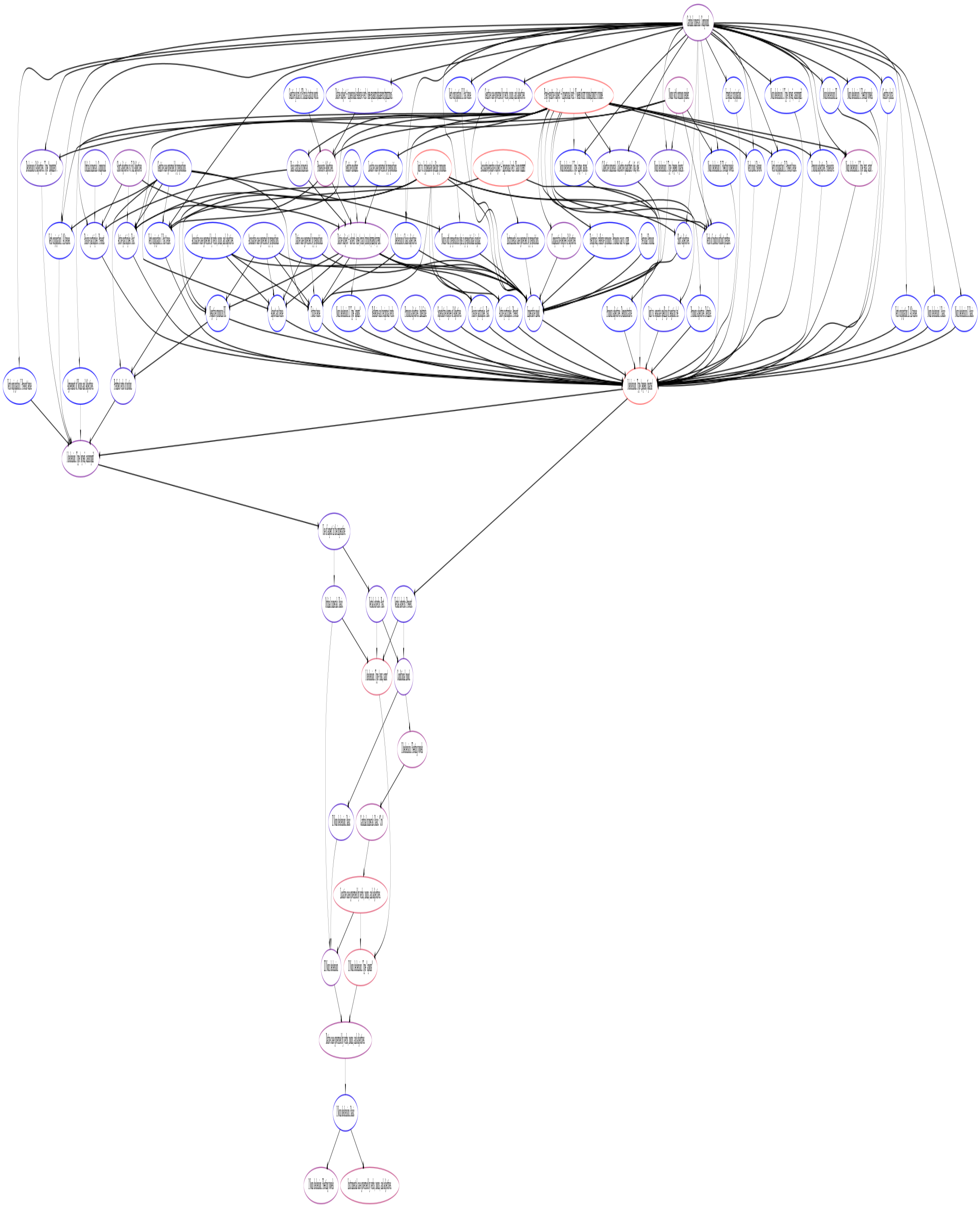
Cardinal numerals: Compound.
Genitive plural of Pluralia tantum words.
Dative subject + impersonal-reflexive verb: [illegible]
Verb conjugation II. Past tense.
Genitive case governed by verbs, nouns, and adjectives.
Prep+genitive subject + impersonal verb: [illegible]
Nouns with common gender.
Irregular conjugation.
Noun declension I. Type [illegible]
Noun declension III.
Noun declension I. Fleeting vowels.
Genitive plural.
Declension of Adjectives. Type [illegible]
Ordinal numerals: Compound.
Short adjectives vs Full adjectives
Genitive case governed by prepositions.
Basic cardinal numerals.
Possessive Adjectives.
Genitive modifier.
Locative case governed by prepositions.
Joint vs. hyphenated spelling: pronouns.
Accusative/genitive subject + impersonal verb: [illegible]
Noun declension I. Type [illegible]
Collective numerals. Collective quantifiers: [illegible]
Noun declension I. Type [illegible]
Noun declension II. Fleeting vowels.
Verb mood. Review.
Verb conjugation II. Present tense.
Pronoun adjectives. Possessive.
Noun declension I. Type [illegible]
Verb conjugation I. All tenses.
Passive participles: Present.
Active participles: Past.
Verb conjugation I. Past tense.
Accusative case governed by verbs, nouns, and adjectives.
Accusative case governed by prepositions.
Dative case governed by prepositions.
Dative subject + adverb: [illegible]
Declension of basic adjectives.
Nouns with prepositions [illegible] in prepositional singular.
Instrumental case governed by prepositions.
Comparative degree of Adjectives.
Reciprocal, reflexive pronouns. Pronouns [illegible]
Personal Pronoun.
Short adjectives.
Verbs of motion without prefixes.
Negative pronouns II.
Aspect and tense.
Future tense.
Noun declension II. Type [illegible]
Reflexive and reciprocal verbs.
Pronoun adjectives. Indefinite.
Superlative degree of Adjectives.
Passive participles: Past.
Active participles: Present.
Imperative mood.
Pronoun adjectives. Demonstrative.
Joint vs. separate spelling of negation [illegible]
Pronoun adjectives. Definite.
Verb conjugation II. All tenses.
Noun declension I. Basic.
Noun declension II. Basic.
Verb conjugation I. Present tense.
Agreement of Nouns and Adjectives.
Prefixed verbs of motion.
I declension. Type [illegible]
I declension. Type [illegible]
Use of aspect in the imperative.
Ordinal numerals: Basic.
Verbal adverbs: Past.
Verbal adverbs: Present.
I declension. Type [illegible]
Conditional mood.
II declension. Fleeting vowels
II Noun declension. Basic
Cardinal numerals: Basic. [illegible]
Locative case governed by verbs, nouns, and adjectives.
III Noun declension
III Noun declension. Type [illegible]
Dative case governed by verbs, nouns, and adjectives.
I Noun declension. Basic
I Noun declension. Fleeting vowels
Instrumental case governed by verbs, nouns, and adjectives.

Figure 11: Prerequisite graph constructed using the threshold-based method.

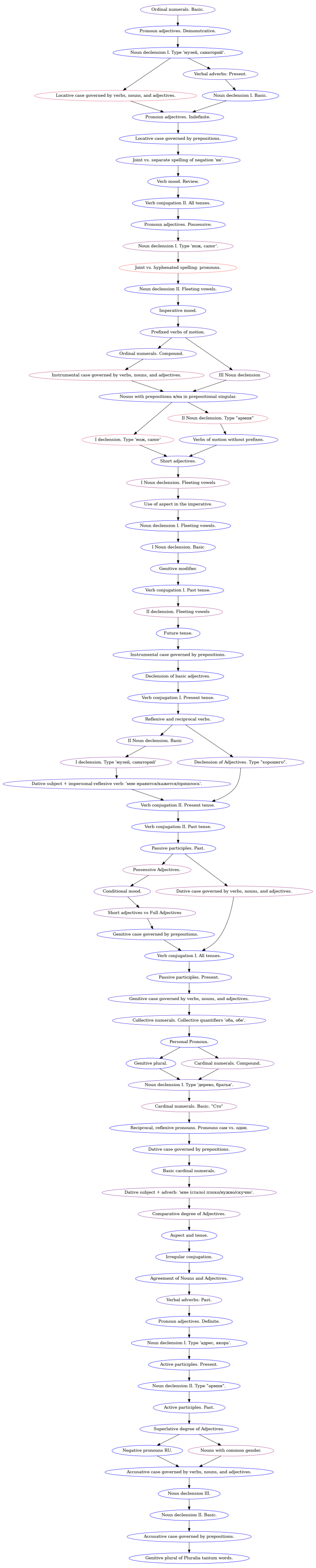
Ordinal numerals. Basic.
Pronoun adjectives. Demonstrative.
Noun declension I. Type 'музей, санаторий'.
Verbal adverbs: Present.
Locative case governed by verbs, nouns, and adjectives.
Noun declension I. Basic.
Pronoun adjectives. Indefinite.
Locative case governed by prepositions.
Joint vs. separate spelling of negation 'не'.
Verb mood. Review.
Verb conjugation II. All tenses.
Pronoun adjectives. Possessive.
Noun declension I. Type 'нож, сапог'.
Joint vs. hyphenated spelling: pronouns.
Noun declension II. Fleeting vowels.
Imperative mood.
Prefixed verbs of motion.
Ordinal numerals. Compound.
Instrumental case governed by verbs, nouns, and adjectives.
III Noun declension
Nouns with prepositions в/на in prepositional singular.
II Noun declension. Type "армия"
I declension. Type 'нож, сапог'
Verbs of motion without prefixes.
Short adjectives.
I Noun declension. Fleeting vowels
Use of aspect in the imperative.
Noun declension I. Fleeting vowels.
I Noun declension. Basic
Genitive modifier.
Verb conjugation I. Past tense.
II declension. Fleeting vowels
Future tense.
Instrumental case governed by prepositions.
Declension of basic adjectives.
Verb conjugation I. Present tense.
Reflexive and reciprocal verbs.
II Noun declension. Basic
I declension. Type 'музей, санаторий'
Declension of Adjectives. Type "хорошего".
Dative subject + impersonal-reflexive verb: 'мне нравится/кажется/пришлось'.
Verb conjugation II. Present tense.
Verb conjugation II. Past tense.
Passive participles. Past.
Possessive Adjectives.
Conditional mood.
Dative case governed by verbs, nouns, and adjectives.
Short adjectives vs Full Adjectives
Genitive case governed by prepositions.
Verb conjugation I. All tenses.
Passive participles. Present.
Genitive case governed by verbs, nouns, and adjectives.
Collective numerals. Collective quantifiers 'оба, обе'.
Personal Pronoun.
Genitive plural.
Cardinal numerals. Compound.
Noun declension I. Type 'дерево, братья'.
Cardinal numerals. Basic. "Сто"
Reciprocal, reflexive pronouns. Pronouns сам vs. один.
Dative case governed by prepositions.
Basic cardinal numerals.
Dative subject + adverb: 'мне (стало) плохо/нужно/скучно'.
Comparative degree of Adjectives.
Aspect and tense.
Irregular conjugation.
Agreement of Nouns and Adjectives.
Verbal adverbs: Past.
Pronoun adjectives. Definite.
Noun declension I. Type 'адрес, якорь'.
Active participles. Present.
Noun declension II. Type "армия".
Active participles. Past.
Superlative degree of Adjectives.
Negative pronouns RU.
Nouns with common gender.
Accusative case governed by verbs, nouns, and adjectives.
Noun declension III.
Noun declension II. Basic.
Accusative case governed by prepositions.
Genitive plural of Pluralia tantum words.

Figure 12: Prerequisite graph constructed using the Grow-Shrink Markov Blanket method.