International Scientific Conference 20 UNITECH 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

LEXICAL AND STRUCTURAL CORRELATES OF ANNOTATOR DISAGREEMENT IN SERBIAN CROSS-LEVEL SEMANTIC SIMILARITY JUDGMENTS

Dragana Radosavljević [1*], Stefan Pitulić [1], Siniša Ilić[1]

[1]Faculty of Technical Sciences, Department of Computer Science and Informatics, University of Priština in Kosovska Mitrovica, Serbia
**Corresponding author: dragana.radosavljevic@pr.ac.rs*

Abstract
Mean semantic similarity ratings can mask disagreement among annotators. We analyze item-level disagreement in 2,000 phrase–sentence and sentence–paragraph pairs from the Serbian CLSS.news.sr dataset, each rated by five annotators on a 0–4 scale. Disagreement is quantified as the normalized mean pairwise absolute difference between ratings. We examine associations with interpretable lexical and structural features and evaluate prediction using repeated 10-fold cross-validation. After accounting for mean-score level as a structural control, the four retained text features increased in-sample R^2 by only 0.00338, while pair type added 0.000054; character-level similarity showed a small conditional association. In prediction, the combined feature model (M2), fitted by ordinary least squares, reduced mean absolute error (MAE) by 0.98% relative to the Jaccard baseline but had higher MAE than the fold-wise constant-median baseline (0.10202 vs. 0.09754). In an MAE-aligned median-regression sensitivity analysis, all feature models achieved the same MAE as the median baseline (0.09754). The out-of-sample R^2 of the combined model remained very low ($R^2_{OOS} = 0.00764$). Thus, the tested feature-based models did not achieve lower MAE than the fold-wise constant-median baseline in this corpus.

Keywords: cross-level semantic similarity; annotator disagreement; semantic textual similarity; human label variation; Serbian language.

1. INTRODUCTION

Semantic Textual Similarity (STS) indicates the extent to which two linguistic units have the same or similar meaning. Most research in this area focuses on comparing units at the same level, in most cases at the level of sentences. Cross-Level Semantic Similarity (CLSS) extends this problem to the comparison of linguistic units at different levels of granularity, such as word–phrase, phrase–sentence, and sentence–paragraph pairs. CLSS was systematically introduced as a separate task within SemEval-2014 Task 3 [1].

Batanović and Miličević Petrović [2] developed CLSS.news.sr, the first publicly available CLSS dataset for a language other than English. The dataset contains 1,000 phrase–sentence pairs and 1,000 sentence–paragraph pairs from the Serbian news domain. Each pair was rated by five annotators on a scale from 0 to 4. The original study primarily focused on dataset construction and validation, preliminary linguistic analysis, and semantic similarity modeling [2]. Subsequent research on Serbian CLSS resources also included a comparison of the news content with the software code comments [3].

A common practice in STS and CLSS tasks is to aggregate individual annotator ratings into a single mean value, which is then used as the reference score. This procedure simplifies subsequent analysis and modeling, but information about the extent to which annotators mutually agree is lost. Thus, when two pairs may have the same mean rating, there is an uncertainty whether that value

This is an open access article licensed under Creative Commons Attribution 4.0 International
Print ISSN: 1313-230X, Online ISSN: 3033-1404

results from almost complete annotator agreement, or it results from substantially different individual judgments.

Recent research on human label variation shows that discrepancies of values obtained by different annotators should not always be treated as the random noise. They may result from different but acceptable interpretations, the subjectivity of the task itself, or other systematic sources of variation [4]. Mostafazadeh Davani et al. [5] likewise show that aggregating annotations can conceal regularities present in individual annotator judgments.

A similar problem arises in semantic similarity assessment. Wang et al. [6] showed that a mean STS score may obscure the structure of individual human judgments. Benito-Santos and Ghajari [7] took it a step further and showed that a disagreement-aware approach enables modeling of the distribution of annotator ratings instead of reducing all judgments to a single mean value.

The study methodologically closest to the present research is that of Zhang and Çöltekin [8], which examines the extent to which item-level variation in graded human ratings can be predicted from text characteristics. Their analysis, however, concerns ratings of offensiveness, toxicity, and hate speech characteristics rather than cross-level semantic similarity.

To the best of our knowledge, no previous study has used the five individual ratings in CLSS.news.sr to quantify item-level annotator disagreement and relates it to interpretable lexical and structural characteristics of cross-level pairs. A further methodological issue arises from the bounded 0–4 rating scale: the attainable level of disagreement depends on the position of the mean rating on the scale. Consequently, higher observed disagreement cannot be interpreted directly as greater semantic ambiguity, and the structural constraints imposed by the rating scale must also be considered.

Accordingly, the main research question of this study was defined as follows: which lexical and structural characteristics of the pairs are associated with the level of annotator disagreement in Serbian cross-level semantic similarity assessment when pair type and the structural constraints of the rating scale are considered?

In addition to the main research question, the study examines whether cross-level pair type provides additional information about annotator disagreement after the score-level constraints and selected textual characteristics are considered. It is also examined the relation between characteristics of the text and annotators disagreement in pairs of the same dataset which are not used for the model estimation.

Three hypotheses were tested. The H1 assumes that the combined feature-based model achieves a lower out-of-sample mean absolute error (MAE) than the simple Jaccard baseline model. The H2 assumes that cross-level pair type provides additional information about annotator disagreement even after the position of the mean rating on the scale and the analyzed characteristics are considered. The H3 assumes that character-level similarity provides additional information beyond the retained word-level and structural characteristics.

The contribution of this paper can be viewed from three perspectives. First, the five individual ratings in CLSS.news.sr are analyzed at the level of individual text pairs rather than only through their mean score. Second, the analysis firstly controls for the position of the mean rating on the bounded scale and then quantifies the correlation between interpretable lexical and structural characteristics with annotator disagreement. Finally, the examinations are conducted to estimate how these characteristics may predict disagreements from the same dataset using repeated cross-validation with fold-specific feature construction to prevent information leakage from the test set into the training process.

The main limitation of the study is that the analysis was conducted on a single CLSS dataset from the news domain and on a limited number of interpretable surface text

features. The results should therefore be viewed primarily as an analysis of regularities present in CLSS.news.sr, while their broader applicability should be examined in other domains and CLSS datasets.

2. MATERIALS AND METHODS

2.1. Dataset and Disagreement Measure

The analysis included all 2,000 pairs from the CLSS.news.sr dataset [2]. Of these, 1,000 are phrase–sentence pairs and the remaining 1,000 are sentence–paragraph pairs. For each pair i, five individual annotator ratings $r_{i1}, r_{i2}, \ldots, r_{i5} \in \{0,1,2,3,4\}$ are available.

Before the analysis, the data were checked for correctness and completeness. No missing or invalid ratings, empty linguistic units, or duplicate pairs were found. It was also confirmed that, in every case, the final score reported in the dataset corresponds to the mean of the five individual annotator ratings.

As an additional data consistency check, Krippendorff's alpha for ordinal data [9] was recalculated. The resulting values were 0.8985 for phrase–sentence pairs and 0.8971 for sentence–paragraph pairs. These values closely match those reported for CLSS.news.sr in the original study [2], providing an additional check on the consistency of the dataset and its preparation for the present analysis.

For each individual pair, item-level annotator disagreement was defined as the mean absolute difference among all ten possible pairs of ratings from the five annotators:

$$D_i = \frac{1}{10} \sum_{1 \le a < b \le 5} |r_{ia} - r_{ib}| \quad (1)$$

Since each pair is rated by five annotators on a scale from 0 to 4, this measure has a bounded maximum value. The theoretical maximum is 2.4 and is obtained when the ratings are distributed between the two extremes of the scale in a 2:3 or 3:2 ratio (extreme disagreement values from the five annotators).

To make the resulting measure easier to compare and interpret, its normalized value was used as the target variable:

$$D_i^* = \frac{D_i}{2.4} \quad (2)$$

so that: $0 \le D_i^* \le 1$.

A value of $D_i^* = 0$ indicates complete agreement among all five annotators, whereas higher values indicate a greater level of disagreement.

2.2. Textual Characteristics

Before feature extraction, all texts were Unicode-normalized to NFC (Unicode Normalization Form C), converted to lowercase, and tokenized in the R software using quanteda 4.5.0 package. Word tokenization was performed with `quanteda::tokens()` using word-level tokenization; punctuation and symbols were removed, numbers and URLs were retained, and hyphenated forms were split. Stop words were retained, and neither stemming nor lemmatization was applied.

For both word- and character-level representations, TF-IDF weights were computed using raw term counts for term frequency (`scheme_tf="count"`) and inverse document frequency (`scheme_df="inverse"`) using logarithm base 10. No additional IDF smoothing, frequency-based feature pruning, or vocabulary-size limit was applied. Character-level features consisted of 3–5-character n-grams generated within individual word tokens, so n-grams did not cross word boundaries.

The following characteristics were calculated for each analyzed pair:

- Jaccard similarity

$$J_i = \frac{|A_i \cap B_i|}{|A_i \cup B_i|,} \quad (3)$$

where A_i and B_i are the sets of unique tokens in the lower-level and higher-level units, respectively.

- Coverage of the lower-level unit

$$C_i = \frac{|A_i \cap B_i|}{|A_i|.} \quad (4)$$

This measure indicates what proportion of the unique tokens in the lower-level unit is also present in the higher-level unit.

- Word TF-IDF cosine similarity

This measure was used to determine pair similarity based on a TF-IDF (Term Frequency–Inverse Document Frequency) representation of words.

- Character 3–5-gram TF-IDF cosine similarity
  For this measure, character n-grams of length 3 to 5 were used. N-grams were formed within individual words and therefore did not cross boundaries between adjacent words.
- Length ratio

$$R_i = \frac{L_i^{low}}{L_i^{high}} \quad (5)$$

where $L_i^{low}$and L_i^{high} stand for number of tokens that belong to the units of lower- and higher- level respectively.

- Log-transformed length of the higher-level text unit

$$L_i^{log} = \log\left(1 + L_i^{high}\right) \quad (6)$$

The terms lower-level unit and higher-level unit are used instead of "shorter" and "longer" text. This terminology is more precise because the level of a linguistic unit is determined by its type rather than solely by the number of tokens. In some examples, the lower-level unit may even contain more tokens than the higher-level unit.

Before fitting the multivariable model, correlations among predictors were examined to avoid simultaneously including several measures carrying very similar information. Jaccard similarity, coverage, and word TF-IDF cosine showed very high mutual correlations $|\rho| > 0.91$, while word TF-IDF cosine was also strongly correlated with character TF-IDF cosine $\rho = 0.812$.

Therefore, coverage was retained in the main model as an interpretable word-level measure, while the character-level measure was retained to test whether it contributes additional information not already captured by word-based measures. Similarity based on n-grams was kept to test whether it impacts the annotator disagreement that was not already present in the words measures. The decision on predictor selection was made based on their mutual correlations before analyzing their relationship with annotator disagreement.

2.3. Statistical Analysis

Both the primary annotator disagreement measure D^* and the mean annotator rating are calculated from the same five individual ratings. At the same time, the possible degree of rating dispersion depends on the position of the mean value on the bounded 0–4 scale. For this reason, score position was introduced into the analysis solely as a structural control rather than as an independent semantic predictor.

Since five integer ratings (assigned by five annotators) are available for each pair, the mean can take 21 possible levels, from 0.0 to 4.0 in increments of 0.2. All levels were sufficiently represented in the dataset, so ScoreLevel was treated as a categorical factor in the primary model. Because both ScoreLevel and D^* are derived from the same five bounded ratings, the R^2 associated with ScoreLevel reflects both the structural dependence between the mean rating and the obtained individual disagreement on the 0–4 scale and empirical differences among the scale levels. It should therefore not be interpreted as the proportion of annotator disagreement caused by the scale constraint.

The relationship between disagreement and the six analyzed characteristics was first examined bivariately using Spearman's correlation coefficient. Bootstrap 95% confidence intervals (CI) were used to assess the reliability of the estimates, while Holm correction was applied for multiple comparisons.

The following OLS (Ordinary Least Squares) linear regression model was used to analyze the predictors:

$$D_i^* = \beta_0 + \beta_1 z(C_i) + \beta_2 z\left(S_i^{char}\right) + \beta_3 z(R_i) + \beta_4 z\left(L_i^{log}\right) + \beta_5 I(PairType_i = SP) + f(ScoreLevel_i) + \varepsilon_i \quad (7)$$

where $f(ScoreLevel_i)$ represents the categorical factor effect of the 21-level mean rating. Here, $z(\cdot)$ denotes z -standardization of the continuous predictors. The indicator

$I(PairType_i = SP)$ equals 1 for sentence–paragraph pairs and 0 for phrase–sentence pairs. Thus, the phrase–sentence pairs represent the reference category and sentence–paragraph pairs the comparison category.

All continuous predictors were standardized before modeling. HC3 (heteroskedasticity-consistent, type 3) robust standard errors were used for statistical inference, while potential multicollinearity among predictors was assessed using VIF/GVIF (Variance Inflation Factor/Generalized Variance Inflation Factor) diagnostics.

Since the D^* is bounded from 0 to 1 interval, and discrete, OLS is used here as an interpretable linear approximation for comparing the text characteristics and level of disagreement; HC3 addresses heteroskedasticity but does not model the bounded outcome distribution itself.

To estimate the incremental contribution of the four retained textual characteristics beyond mean-score position, two models were first compared. The first contained only ScoreLevel as a structural control, whereas the second additionally included coverage, character-level similarity, length ratio, and log-transformed higher-level unit length. Pair type was subsequently added in a third model to quantify its additional contribution to disagreement when scale position and textual characteristics are considered.

Additional contribution to the text characteristics is calculated as the difference of the deterministic coefficients:

$$\Delta R^2_{text} = R^2_{ScoreLevel+text} - R^2_{ScoreLevel} \quad (8)$$

The ΔR^2_{text} value thus shows how much R^2 increases on the analyzed sample when four text features are added to the model containing only the average grade level. In the same way, the change in R^2 after including the pair type is denoted as ΔR^2_{pair}. These values are used to estimate the additional association of certain groups of predictors with disagreement and are not interpreted as indicators of cause-and-effect relationships.

The stability of ΔR^2_{text} was assessed using 3,000 bootstrap resamples at the individual text-pair level, and 95% percentile confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. As sensitivity checks, ScoreLevel was alternatively modeled using a natural spline with four degrees of freedom, and the analysis was repeated using pooled ordinal item-level disagreement as an alternative outcome.

Specifically, this alternative outcome was defined as the item-specific observed disagreement computed using Krippendorff's ordinal distance, normalized by the pooled expected ordinal disagreement estimated across all 2,000 pairs [7,9].

As a further sensitivity check, the ScoreLevel-only model was refitted after excluding observations at the deterministic endpoint score levels of 0.0 and 4.0. Since all five annotators estimate the same at the endpoint values, the disagreement is zero, and the estimation results between ScoreLevel and the disagreements may be validated.

2.4. Predictive Evaluation

Predictive performance on held-out pairs from the same dataset was evaluated using repeated 10-fold cross-validation with five repetitions. Mean absolute error (MAE) was prespecified as the primary evaluation metric. In addition to the mean baseline, a fold-wise median baseline was included as a stricter reference because the median minimizes absolute error.

Five models were compared. M1, M2-no-char, and M2 were fitted as ordinary least-squares regression models within each training fold.

- M0-mean – a constant baseline predicting the mean target value of the corresponding training fold for all held-out pairs;
- M0-median – a constant baseline predicting the median target value of the corresponding training fold for all held-out pairs;
- M1 – an OLS baseline model using Jaccard similarity as the sole predictor;

- M2-no-char – a model including coverage, length ratio, log-transformed higher-level unit length, and pair type;
- M2 – an extension of M2-no-char with character TF-IDF cosine similarity.

ScoreLevel, the mean annotator rating, and any other information directly or indirectly derived from human ratings were not included in the predictive models. Fold assignment was stratified by pair type and quintiles of D^* to improve balance across folds. D^* was used as the response variable and for fold stratification, but was not included among the predictors. Only training-fold D^* values were used to fit each model. Special care was taken to ensure that text pairs containing the same lower- or higher-level normalized unit were not distributed between the training set and the test set. Such pairs were classified in the same part of the data in each repetition. Duplicate screening identified no repeated normalized lower-level units and one higher-level unit shared by two pairs.

As an additional sensitivity analysis, potential near-duplicate text units were identified from normalized text using character 3–5-gram TF-IDF cosine similarity. Pairs were treated as related when any of their lower- or higher-level units had cosine similarity of at least 0.95, and all pairs belonging to the same connected component were constrained to the same fold. A more permissive threshold of 0.90 was examined as an additional robustness check. The two thresholds were specified before examining the sensitivity-analysis results and were not selected based on predictive performance. Near-duplicate screening used textual information only and did not use D^* or any other information derived from annotator ratings.

To prevent information transfer from the validation set to the training set, a dictionary of character n-grams and the corresponding IDF weights were determined separately in each training set. Thus, formed dictionary and weights were then applied to the appropriate extracted part of the data. No predictor standardization was applied in the predictive models.

Because the primary predictive models were fitted by ordinary least squares, which minimizes squared rather than absolute error, an additional sensitivity analysis was performed using median quantile regression ($\tau = 0.5$). This analysis assessed whether the predictive conclusions were retained when model fitting was aligned with the MAE metric that was selected as the primary error metric. Three quantile-regression (QR) models were evaluated at $\tau = 0.5$: QR-M1 using Jaccard similarity, QR-M2-no-char using coverage, length ratio, log-transformed higher-level unit length, and pair type, and QR-M2 additionally including character TF-IDF cosine similarity. The same group-aware cross-validation folds and fold-specific character vocabulary and IDF weights as in the primary predictive analysis were used. This analysis was treated as a sensitivity check and did not replace the primary OLS analysis.

MAE was prespecified as the primary metric. Root mean squared error (RMSE), out-of-sample R^2, and Spearman's correlation coefficient were reported as complementary measures. Out-of-sample R^2 was defined relative to the observed mean within each cross-validation repetition. Each pair received one out-of-fold prediction per repetition; metrics were calculated separately for each repetition and then averaged. For paired model comparisons, pair-level errors were first calculated and then averaged across the five repetitions. To compare models, the errors between text pairs were averaged from five repetiotions and then resampled with replacement using 3,000 paired bootstrap resamples, and 95% percentile confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

All analyses were performed in R 4.6.0; random seed 20260828 was used.

3. RESULTS

3.1. Item-Level Disagreement

Across all 2,000 analyzed pairs, the mean item-level disagreement was 0.1454, with a standard deviation of 0.1231 and a median of

0.1667. For 713 pairs, or 35.65% of the full dataset, all five annotators scored equally all pairs, meaning that more than one third of the examples showed no disagreement at all.

By pair type, mean disagreement was almost identical in the two groups. For phrase–sentence pairs the mean was 0.1442, whereas for sentence–paragraph pairs it was 0.1466. The similarity of these values indicates that, at the level of basic descriptive statistics, the two cross-level pair types show little difference in the mean annotator disagreement.

Table 1. Basic characteristics of item-level disagreement

Pair type	Phrase–sentence	Sentence–paragraph	Overall
N	1000	1000	2000
Mean D^*	0.1442	0.1466	0.1454
Std. dev.	0.1252	0.1210	0.1231
Median	0.1667	0.1667	0.1667
Complete agreement	36.90%	34.40%	35.65%
Max D^*	0.5833	0.7500	0.7500
Global ordinal Krippendorff's α	0.8985	0.8971	-

Bivariate analysis showed weak positive associations between disagreement and the lexical similarity measures. Character 3–5-gram TF-IDF cosine similarity showed the strongest association ($\rho = 0.169$, 95% CI $[0.124, 0.212]$, $p_{Holm} < 0.001$), followed by word TF-IDF cosine ($\rho = 0.136$, 95% CI $[0.090, 0.179]$, $p_{Holm} < 0.001$), Jaccard similarity ($\rho = 0.121$, 95% CI $[0.078, 0.167]$, $p_{Holm} < 0.001$), and coverage ($\rho = 0.118$, 95% CI $[0.073, 0.160]$, $p_{Holm} < 0.001$). Length ratio ($\rho = 0.003$, 95% CI $[-0.039, 0.047]$) and log-transformed higher-level length ($\rho = 0.011$, 95% CI $[-0.036, 0.055]$) showed practically no associations (both $p_{Holm} = 1.000$).

The character-level measure had been retained for the further analysis before examining its association with disagreement. This decision was based on predictor redundancy and its conceptually distinct content relative to the word-level measures.

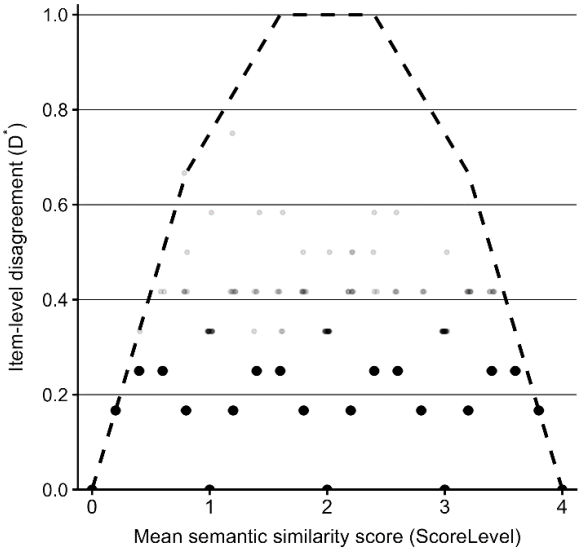


Fig. 1. *Item-level annotator disagreement across mean semantic similarity scores. Individual points represent text pairs, solid points denote observed medians, and the dashed line indicates the theoretical maximum D^* permitted by five ratings on the 0–4 scale.*

Figure 1 shows the structural constraint imposed by the 0–4 scale: the theoretically attainable maximum D^* depends on the position of the mean rating; the largest disagreement is at the middle,and is substantially lower near the endpoints of the scale.

3.2. Multivariable Analysis Controlling for Score Level and Other Predictors

After score position and the other retained characteristics were included in the model, character-level similarity was the only textual characteristic showing a clear conditional association with annotators' disagreement ($\beta = -0.00936$, 95% CI $[-0.01604, -0.00268]$, $p = 0.006$). The negative sign of the coefficient shows that (at the same values of the other predictors) greater similarity at the character level was associated with a slightly lower degree of disagreement.

No clear association with disagreement was found for coverage, length ratio, or higher-level unit length after the other predictors were considered.

Cross-level pair type likewise did not show a clear conditional association ($\beta = 0.00424$, 95% CI $[-0.01039, 0.01888]$, $p = 0.570$). Accordingly, H2 was not supported.

Table 2. Key coefficients of the main OLS model with HC3 robust standard errors

Predictor	Coefficient β	95% CI	p
Coverage	-0.00338	[-0.00793, 0.00118]	0.146
Character similarity	**-0.00936**	**[-0.01604, -0.00268]**	**0.006**
Length ratio	0.00135	[-0.00303, 0.00573]	0.545
Log higher-level length	-0.00114	[-0.00921, 0.00693]	0.782
PairType	0.00424	[-0.01039, 0.01888]	0.570

Note: continuous predictors were standardized. The coefficient for PairType represents the difference for sentence–paragraph pairs relative to the reference category of phrase–sentence pairs.

Adjusted $GVIF^{1/(2df)}$ values ranged from 1.03 to 2.41, indicating that there was no pronounced multicollinearity problem after the prior reduction of predictors.

The ScoreLevel-only model had $R^2 = 0.6414$. Adding the four retained textual characteristics increased R^2 to approximately 0.6447. The corresponding increase was $\Delta R^2_{\text{text}} = 0.00338$, with a 95% bootstrap confidence interval of [0.00152,0.00751]. Thus, the four textual characteristics increased the total in-sample R^2 by approximately 0.338 percentage points. Relative to the variance remaining after ScoreLevel was taken into account, the corresponding partial R^2 was approximately 0.00942. Adding PairType subsequently increased R^2 by only $\Delta R^2_{\text{pair}} = 0.000054$, yielding a full-model R^2 of approximately 0.6448.

Adding character-level similarity specifically increased R^2 by approximately 0.00149. In the primary model specification, this result provides limited support for H3, indicating a small incremental contribution of the character-level measure beyond the other retained characteristics. However, the magnitude of this additional contribution is small.

The negative character-level coefficient was retained in both specification sensitivity analyses. The coefficient remained negative when ScoreLevel was modeled with a spline ($\beta = -0.01343$, 95% CI $[-0.02340, -0.00345]$, $p = 0.008$) and when pooled ordinal item-level disagreement was used as the alternative outcome ($\beta = -0.01031$, 95% CI $[-0.01796, -0.00267]$, $p = 0.008$).

As a further sensitivity check, observations at the deterministic endpoint score levels of 0.0 and 4.0 ($n = 391$) were excluded. At these points all five annotators agree, so disagreement is equal to zero. In the reduced sample, the ScoreLevel-only model yielded $R^2 = 0.4574$, compared with $R^2 = 0.6414$ in the full sample. Thus, a substantial in-sample association between score position and disagreement remained after the deterministic endpoint levels were removed. Because the two R^2 values were obtained from different samples, their difference should not be interpreted as a decomposition of the endpoint contribution.

3.3. Predictive Evaluation

The results of the group-aware repeated cross-validation are presented in Table 3. The fold-wise constant-median baseline model based on median values achieved the lowest MAE overall (0.09754), whereas M2 achieved the lowest RMSE (0.12260) and the highest out-of-sample R^2 (0.00764) and Spearman correlation (ρ = 0.128). Additional sensitivity analyses were performed on very similar textual units using character-level cosine thresholds of 0.95 and 0.90; these analyses are reported below.

Table 3. Out-of-sample model performance

Model	MAE	RMSE	R^2_{oos}	Spearman
M0-median	**0.09754**	0.12490	-0.02993	–
M0-mean	0.10365	0.12308	-0.00005	-0.005
M1 - Jaccard	0.10303	0.12285	0.00368	0.116
M2-no-char	0.10263	0.12279	0.00458	0.102
M2	0.10202	**0.12260**	**0.00764**	**0.128**

Spearman correlation coefficient could not be calculated for M0-median because its out-of-fold predictions are constant. The paired MAE difference between the constant-median baseline and M2 was $MAE_{median} -$

$MAE_{M2} = -0.00448$ (95% CI $[-0.00561, -0.00338]$); In other words M2 had a 4.59% higher MAE. Conversely, $RMSE_{median} - RMSE_{M2} = 0.00230$ (95% CI $[0.00109, 0.00350]$), indicating lower RMSE for M2.

Compared with the Jaccard baseline M1, M2 reduced MAE by 0.001005 (95% CI $[0.000615, 0.001381]$), corresponding to a relative reduction of 0.98%. Because H1 specifically compares M2 with the Jaccard baseline using the prespecified primary MAE metric, the primary analysis supports H1 in this narrow comparison. The robustness of this result to an MAE-aligned estimation method is examined below. The corresponding RMSE difference was 0.000244 (95% CI $[-0.000137, 0.000640]$), so a clear RMSE advantage was not established.

Character-level similarity also provided a small predictive increment. The difference $MAE_{M2-no-char} - MAE_{M2} = 0.000607$ (95% CI $[0.000260, 0.000966]$) corresponds to a relative MAE reduction of approximately 0.59%, providing limited predictive support for H3. The corresponding RMSE difference, $RMSE_{M2-no-char} - RMSE_{M2} = 0.000189$ (95% CI $[-0.000178, 0.000546]$), and difference between the models was not clearly identified.

As an MAE-aligned sensitivity analysis, the same feature *sets* were fitted using median quantile regression ($\tau = 0.5$). M0-median, QR-M1, QR-M2-no-char, and QR-M2 all achieved essentially identical predictive performance (MAE=0.09754, RMSE=0.12490, and $R^2_{OOS} = -0.02993$). Thus, neither the combined textual feature set nor character-level similarity improved MAE under median regression. The small OLS MAE advantage of M2 over the Jaccard baseline was'nt therefore confirmed when the applied estimation procedure was directly aligned with the mean absolute error, and the predictive contribution of character-level similarity was likewise not reproduced in this sensitivity analysis.

Despite these small pair-specific gains, overall predictive power remained very low ($R^2_{OOS} = 0.00764$; *Spearman* $\rho = 0.128$ for M2). Thus, the analyzed textual characteristics contain a weak pair-specific signal but do not outperform the fold-wise constant-median baseline under the prespecified primary MAE metric.

Additional calculations when very similar units were considered, did not change the qualitative conclusions of the OLS comparisons. At the 0.95 threshold, 14 pairs belonged to multi-pair components; H1 remained supported ($\Delta MAE = 0.000974$, 95% CI $[0.000614, 0.001348]$), and the small predictive contribution of character similarity was retained ($\Delta MAE = 0.000620$, 95% CI $[0.000276, 0.000961]$). At the more permissive 0.90 threshold, 20 pairs belonged to multi-pair components, and the corresponding differences remained positive for H1 ($\Delta MAE = 0.000876$, 95% CI $[0.000475, 0.001262]$) and for character similarity ($\Delta MAE = 0.000584$, 95% CI $[0.000237, 0.000941]$). No component contained more than two pairs. The constant-median baseline retained the lowest MAE, while R^2_{OOS} for M2 remained very low (0.00584 and 0.00417 at thresholds 0.95 and 0.90, respectively).

4. DISCUSSION

Aggregation into a single mean score discards item-level dispersion by definition. The present analysis shows that the examined text surface characteristics are associated with only a small additional proportion of this variation. The decision to analyze disagreement at the item level is consistent with disagreement-aware approaches [6,7].

The strongest in-sample structure involved mean-score position. The categorical ScoreLevel control alone reproduced approximately 64.14% of the in-sample variance in D^*. Because ScoreLevel and D^* were calculated from the same five bounded ratings, R^2 combines structural coupling between the mean rating and attainable dispersion with empirical differences across score levels. It must therefore not be interpreted as the proportion of disagreement caused by scale constraints.

Additional analysis showed that this is not the consequence of the extreme values on the scale. After observations with mean-score levels of 0.0 and 4.0 were removed, the ScoreLevel-only model still yielded $R^2 = 0.4574$. Thus, the strong association between score position and disagreement was not confined to the deterministic endpoints, although the full- and reduced-sample R^2 ($R^2 = 0.6414$ and $R^2 = 0.4574$)values should not be interpreted as a variance decomposition.

ScoreLevel was therefore included to estimate associations between textual characteristics and annotators' disagreements conditional on mean-score position. This defines a within-ScoreLevel estimand; it does not isolate a causal effect of the scale.

This distinction is particularly important for character-level similarity. In the bivariate analysis, its association with D^* was positive ($Spearman\ \rho = 0.169$), whereas after controlling for ScoreLevel and the other retained predictors, the coefficient changed direction ($\beta = -0.00936$).

This means that the ratio observed when similarity at the character level is considered separately is not the same as the ratio obtained when other characteristics are simultaneously considered. A negative coefficient in the model describes a situation in which couples with the same position of average grade and comparable values of other predictors are compared. Part of this change may be a consequence of conditioning on average semantic similarity, so a negative coefficient should not be seen as a contradiction to a positive individual correlation.

Conditional on ScoreLevel and the other included predictors, higher character similarity was associated with slightly lower disagreement. The present analysis supports only a conditional association and provides no basis for a causal interpretation.

The predictive analysis further illustrates the limitations of text surface feature-based approaches. In the primary OLS analysis, M2 reduced MAE by approximately 0.98% relative to the Jaccard baseline, providing narrow support for H1 within that specification. However, the fold-wise constant-median baseline achieved the lowest MAE overall (0.09754 compared with 0.10202 for M2). More importantly, when the same textual feature sets were fitted using median quantile regression, which aligns model estimation with the primary MAE criterion, QR-M1, QR-M2-no-char, and QR-M2 all performed essentially identically to the median baseline. Thus, the small model M2 advantage over the Jaccard baseline was not robust to the estimation loss and should not be interpreted as general predictive superiority of the combined feature model. The paired bootstrap confirmed this difference ($M0 - median - M2 = -0.00448$, 95% CI $[-0.00561, -0.00338]$).

On the other hand, M2 achieved a lower RMSE than the median baseline (0.12260 vs. 0.12490) and the highest out-of-sample R^2 and Spearman correlation among the evaluated feature-based models. The RMSE difference between the median baseline and M2 was 0.00230 (95% CI $[0.00109, 0.00350]$). Thus, the preferred model depends on the evaluation metric: the textual characteristics contain a weak within-dataset pair-specific signal, but this signal is not strong enough to outperform a simple median predictor under the prespecified primary MAE metric.

The same qualitative pattern was retained in the very similar textual units sensitivity analyses at cosine thresholds of 0.95 and 0.90, although the already low R^2_{OOS} of M2 decreased to 0.00584 and 0.00417, respectively. This indicates that the presence of a small number of highly similar text units was not the main reason for the observed results.

This result complements the findings of Zhang and Çöltekin [8], who showed that item-level variation in human ratings in other graded annotation tasks can be partially predicted from text properties. Direct comparison of predictive performance is limited because the studies concern different annotation tasks and potentially different sources of annotator disagreement.

The results also did not support H2. Pair type (phrase–sentence vs. sentence–paragraph) did not provide clear additional information after ScoreLevel and the retained characteristics were included in the model. This result does not imply that granularity has no role in disagreement during semantic similarity assessment. It only shows that, in the analyzed dataset, the division into these two types of cross-level pairs does not make a clear contribution to explaining annotator disagreement.

Several limitations should be considered when interpreting the results. The analysis was conducted on a single Serbian CLSS dataset from the news domain and included only two types of cross-level pairs. Five individual ratings are available for each example, limiting detailed analysis of the full distribution of annotator judgments. The publicly available data also do not allow control for the person who initially constructed each pair. In addition, the study deliberately used interpretable text surface features and therefore did not include more complex semantic, syntactic, pragmatic, or discourse relations. Finally, group-aware repeated cross-validation evaluates held-out pairs from the same corpus rather than transfer to new domains, and the highly discrete target distribution limits the interpretation of continuous-regression metrics.

As a result, a large proportion of the variation in annotator disagreement remains unexplained. It may be influenced by more complex semantic relations between the units, different inferences made by annotators, broader context, and individual rating patterns. However, the available data do not permit reliable separation and estimation of the contribution of each of these factors.

5. CONCLUSION

This study analyzed item-level disagreement among five annotators in Serbian cross-level semantic similarity assessment. The examined lexical and structural characteristics showed systematic but small associations with disagreement after conditioning on mean-score position.

Mean-score position was associated with a large proportion of the in-sample variation in D^*, but this reflects both structural coupling between the mean and attainable dispersion and empirical differences across score levels. Adding the four textual characteristics increased R^2 by only 0.00338, while pair type contributed only an additional 0.000054. Character-level similarity contributed a small additional conditional association, whereas pair type did not provide meaningful additional information. The median-regression sensitivity analysis partly addresses the mismatch between MAE evaluation and least-squares fitting, but it does not remove the fundamental limitation imposed by the discrete and highly concentrated target distribution.

M2 achieved a slightly lower MAE than the Jaccard baseline, but the fold-wise constant-median baseline achieved a lower MAE than M2. Moreover, an MAE-aligned median-regression sensitivity analysis yielded essentially identical performance for QR-M1, QR-M2-no-char, QR-M2, and the median baseline (MAE = 0.09754). Thus, the small OLS advantage over the Jaccard baseline was not robust to the estimation loss. Although M2 achieved the lowest RMSE and the highest out-of-sample R^2 and Spearman correlation among the feature-based OLS models, overall pair-specific predictive power remained very low ($R^2_{OOS} = 0.00764$). H1 therefore receives only specification-dependent support only in the narrow framework of the pre-defined comparison of the M2 model and the Jaccard base model according to the MAE, while the predictive component of H3 is not robust across estimation methods. According to the results, H2 was not supported.

The current results do not establish reliable automatic identification of CLSS pairs requiring additional rating or adjudication, because high-disagreement detection was not evaluated as a separate classification or ranking task.

The findings underscore the value of retaining individual annotator ratings rather than representing all annotation information solely through a single aggregated reference value. Therefore, whenever possible, they

should be preserved and analyzed separately, especially in cases where discrepancies between annotators can be a significant part of the phenomenon being studied. Future research should examine richer semantic and linguistic characteristics, larger annotator samples, explicit high-disagreement detection, and validation on other CLSS datasets, domains, and languages.

Author Contributions:

D.R.: Conceptualization, methodology, software, formal analysis, validation, visualization, writing – original draft, writing – review and editing, supervision, project administration.
S.P.: Methodology, validation, formal analysis, writing – review and editing.
S.I.: Validation, writing – review and editing, supervision.
All authors have read and agreed to the final version of the manuscript.

ACKNOWLEDGEMENTS

This research was supported by the Erasmus+ Capacity Building in Higher Education project 101179358-Res-Uni-ERASMUS-EDU-2024-CBHE "Research Universities (Res-Uni)".
The authors would like to thank the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for funding the scientific research work, contract no. 451-03-34/2026-03/200155, realized by the Faculty of Technical Sciences in Kosovska Mitrovica, University of Priština.

The funding bodies had no role in the study design, data analysis, interpretation of the results, preparation of the manuscript, or decision to submit the paper.

REFERENCES

[1] D. Jurgens, M. T. Pilehvar and R. Navigli, "SemEval-2014 Task 3: Cross-Level Semantic Similarity", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 17–26, 2014. doi: 10.3115/v1/S14-2003. https://aclanthology.org/S14-2003.pdf (accessed 30 August 2026).

[2] V. Batanović and M. Miličević Petrović, "Cross-Level Semantic Similarity for Serbian Newswire Texts", Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022), pp. 1691–1699, 2022. https://aclanthology.org/2022.lrec-1.180.pdf (accessed 30 August 2026).

[3] M. Miličević Petrović, V. Batanović, R. Trnavac and B. Kovačević, "Cross-Level Semantic Similarity in Newswire Texts and Software Code Comments: Insights from Serbian Data in the AVANTES Project", Proceedings of JTDH 2022, pp. 124–131, 2022. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Petrovic-et-al_Cross-Level-Semantic-Similarity-in-Newswire-Texts-and-Software-Code.pdf (accessed 30 August 2026).

[4] B. Plank, "The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation", Proceedings of EMNLP 2022, pp. 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731. https://aclanthology.org/2022.emnlp-main.731.pdf (accessed 30 August 2026).

[5] A. Mostafazadeh Davani, M. Díaz and V. Prabhakaran, "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations", Transactions of the Association for Computational Linguistics, vol. 10, pp. 92–110, 2022. doi: 10.1162/tacl_a_00449. https://aclanthology.org/2022.tacl-1.6.pdf (accessed 30 August 2026).

[6] Y. Wang, S. Tao, N. Xie, H. Yang, T. Baldwin and K. Verspoor, "Collective Human Opinions in Semantic Textual Similarity", Transactions of the Association for Computational Linguistics, vol. 11, pp. 997–1013, 2023. doi: 10.1162/tacl_a_00584. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00584/2154521/tacl_a_00584.pdf (accessed 30 August 2026).

[7] A. Benito-Santos and A. Ghajari, "Beyond Averages: Learning with Annotator Disagreement in STS", Proceedings of EMNLP 2025, pp. 35551–35557, 2025. doi: 10.18653/v1/2025.emnlp-main.1800. https://aclanthology.org/2025.emnlp-main.1800.pdf (accessed 30 August 2026).

[8] L. Zhang and Ç. Çöltekin, "Quantifying and Predicting Disagreement in Graded Human Ratings", Proceedings of NLPerspectives 2026, pp. 33–43, 2026. doi: 10.63317/4qy8nuowzhpy. https://lrec.elra.info/lrec2026-ws-nlperspectives-04 (accessed 30 August 2026).

[9] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations", Human Communication Research, vol. 30, no. 3, pp. 411–433, 2004. doi: 10.1111/j.1468-2958.2004.tb00738.x. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x (accessed 30 August 2026).