SimilEx: the First Italian Dataset for Sentence Similarity with Natural Language Explanations

Chiara Alzetta[1], Felice Dell'Orletta[1], Chiara Fazzone[1] and Giulia Venturi[1]

[1] *ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy*

Abstract

Large language models (LLMs) demonstrate great performance in natural language processing and understanding tasks. However, much work remains to enhance their interpretability. Annotated datasets with explanations could be key to addressing this issue, as they enable the development of models that provide human-like explanations for their decisions. In this paper, we introduce the SimilEx dataset, the first Italian dataset reporting human judgments of semantic similarity between pairs of sentences. For a subset of these pairs, the annotators also provided explanations in natural language for the scores assigned. The SimilEx dataset is valuable for exploring the variability in similarity perception between sentences and among human explanations of similarity judgments.

Keywords

Sentence similarity, Italian dataset, human judgements, explanations, annotation

1. Introduction and Motivation

Large language models (LLMs) display impressive linguistic skills and demonstrate outstanding performances on a variety of tasks concerning natural language processing and understanding. This is particularly true for the most recent and ground-breaking models such as GPT-3.5\4 [1], LLama-2 [2] and Gemini [3]. LLMs, however, also present risky limitations such as lack of factuality [4, 5], poor interpretability [6, 7] and hallucinations [8]. Consequently, it has become important to verify whether these models are explainable, and specifically whether they can provide human-like explanations using natural language for decisions made [9, 10]. The ability of LLMs to explain the reasoning needed to solve a given task is fundamental, particularly for tasks where there is no established or shared evaluation protocol or benchmark.

Annotated datasets with explanations are key to addressing this issue, as they enable the development of models that provide human-like explanations for their decisions. Therefore, multiple datasets have been created with free-form explanations to be incorporated into the model training process and used as benchmarks at test time, mostly focusing on English [10]. Some examples are the e-SNLI dataset [11], a version of the Stanford Natural Language Inference (SNLI) dataset [12] enriched with human-annotated explanations, and the Common Sense Explanations (CoS-E) [13] and Semi-Structured Explanations for COPA (COPA-SSE) [14] datasets, which include natural language explanations for commonsense reasoning. To the best of our knowledge, the only existing dataset enriched with explanations for Italian is 'e-RTE-3-it' [15], an Italian version of the RTE-3 dataset for textual entailment.

In this paper, we introduce the SimilEx dataset[1], as far as we are aware, the first Italian dataset of 2,112 pairs of sentences manually annotated for semantic similarity. About half of the pairs are further enriched with free-form human-written explanations that justify the similarity score.

The identification of textual similarity is a natural language understanding (NLU) task that involves determining the degree of semantic equivalence between two texts [16, 17]. It is a foundational NLU problem relevant to many applications such as summarisation, question answering and conversational systems [18]. Despite its relevance, this task is highly challenging even for humans due to its subjective nature: human annotations often widely disagree on similarity scores [19] suggesting that the cues driving sentence similarity are neither well codified nor transparent and that their perceived relevance may vary among annotators. Possibly due to these challenges, and as far as we know, datasets including human explanations for the sentence similarity task are lacking. However, they are invaluable as they force annotators to reason about their choices and identify the most relevant traits influencing their annotations.

Contributions. In this paper, we *i)* introduce SimilEx, the first Italian dataset featuring human annotations and explanations of sentence semantic similarity; *ii)* provide an extensive study of the degree of subjectivity in the perception of sentence semantic similarity; and *iii)* investigate the relationship between the stylistic variation of

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

chiara.alzetta@ilc.cnr.it (C. Alzetta); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); chiara.fazzone@ilc.cnr.it (C. Fazzone); giulia.venturi@ilc.cnr.it (G. Venturi)

[1] The dataset is freely available at http://www.italianlp.it/resources/.

the paired sentences and the human ratings and natural language explanations of sentence semantic similarity.

2. The SimilEx Dataset

2.1. Data Collection

The sentence pairs of SimilEx are acquired from a collection of novels from the late XIX century translated into Italian. We used Sentence-BERT (SBERT) [20] to combine pairs of sentences to present to annotators. SBERT is a modification of BERT [21] made adequate to produce sentence embeddings that can be easily compared to evaluate their similarity using cosine similarity, which ranges from 0 (no similarity) to 1 (identical sentences). We included in the SimilEx dataset only pairs obtaining a similarity score ≥ 0.65, for a total of 2,112 sentence pairs.

The textual genre of the sentences (i.e., novels) introduces specific stylistic properties that cause potential differences from standard Italian. We assessed the linguistic style of SimilEx sentences using Profiling-UD [22][2], a web-based tool that captures multiple aspects of sentence structure. The tool extracts around 130 properties representative of the underlying linguistic structure of a sentence, derived from raw, morphosyntactic, and syntactic levels of sentence annotation, all based on the Universal Dependencies (UD) formalism [23]. These properties have been shown to be highly predictive when used as features by learning models in various classification tasks, such a Automatic Readability and Linguistic Complexity Assessment or Native Language Identification. Among these caracteristics, the average length computed on Similex sentences is 30.18 tokens (±22.36), above the average length of standard Italian sentences, typically around 20 tokens. Interestingly, within pairs, the average length difference is 17.02 tokens (±19.55). This value, combined with such a high standard deviation, suggests a large variability of style within the pairs. This notable variability extends, e.g., to the distribution of subordinate clauses and lexical overlap. Within pairs, the average difference in the number of subordinate clauses is 2.25 (±1.81), and the overlap of content words is 12.60%, which are significant given that this variation occurs within individual sentence pairs. Having pairs with such stylistic differences provides an opportunity to investigate the impact of stylistic variation on the perception of similarity.

2.2. Human Similarity Annotation

Sentence pairs of SimilEx were annotated through the online crowdsourcing platform Prolific[3]. Annotators were recruited among native Italian speakers and presented with a questionnaire of 30 pairs plus 2 control pairs.

Annotation Guidelines. The task consisted of scoring each sentence pair of the questionnaire for the perceived sentence similarity using a 5-point Likert scale, where 1 is described as *"Completamente diverse"* (Completely different) and 5 as *"Pressoché identiche"* (Almost identical). Any formal definition of similarity is provided, only a few examples of highly similar and highly different pairs along with motivations for the extreme similarity scores, as shown in the annotation instructions provided to the annotators fully reported in Appendix C. This represents the main novelty of our approach compared to the methodology used to create datasets for Semantic Textual Similarity tasks, typically organized within the SemEval evaluation campaign (see among the others [24, 18]). These datasets are usually built with clear and specific instructions for annotators, who are explicitly asked to evaluate whether paired text portions refer to the same person, action, or event, or to focus their judgment on similarity types such as the same author, time period, or location. Some examples of annotation with similarity scores averaged across annotators are shown in Table 1.

Demographics. Participants could share information about their age, gender and occupation and complete multiple questionnaires. Eventually, 317 distinct participants took part in the study. After a preliminary analysis, we excluded 34 annotators deemed unreliable because they either took too short to complete the questionnaire, assigned systematically divergent scores compared to the rest of the participants, failed the control questions or submitted blank answers. The resulting dataset includes 2,112 sentence pairs annotated by the remaining 283 annotators, who took 18 minutes on average to complete a questionnaire[4]. Each pair received a minimum of 5 and a maximum of 7 annotations from different participants. The set of annotators is quite balanced for gender (51% males) and the average age of annotators is 27.05 (±6.56). Regarding occupation, 50% of participants indicated that they have a full- or part-time job, around 25% declared themselves unemployed, and the remaining 25% preferred not to disclose their occupational status.

2.3. Human Explanations of Similarity

We recruited 2 native Italian speakers who volunteered to enrich the pairs of sentences with free-form explanations. These annotators are graduate students, one male and one female, aged 23 years. They were asked to score the similarity of a random subset of 907 sentence pairs on the same 5-point Likert scale as the other participants. Additionally, they should provide a short explanation for

[2] The complete set of linguistic characteristics used for the stylistic analysis can be found in Appendix B.

[3] https://www.prolific.com/

[4] The compensation is fair according to the platform: 6.30£/hour.

Sentence 1	Sentence 2	Mean Similarity Score
Sì, grazie a Dio non è male.	*Io invece non ce l'ho: tante grazie!*	1.1
Non hanno mandato a prendere il latte fresco?	*E per me, chiedi almeno del latte.*	2.6
Solo lo zar può far la grazia.	*Voglio chiedere la grazia allo zar.*	3.1
Accidenti a voi, mi fate perdere il filo!	*Intanto voi però mi avete fatto perdere il filo.*	4.7

Table 1
Pairs of sentences annotated with similarity scores averaged across annotators.

their scores, in the form of a single concise sentence.

3. Human Similarity Perception

The first analysis of SimilEx focuses on the exploration of the similarity judgments expressed by annotators using the scores. Note that for this analysis all scores were considered, including those of the two students who provided the explanations. Firstly, we computed the Pearson correlation between the average similarity scores of sentence pairs and SBERT scores, obtaining $r = 0.28$ ($p < 0.001$). This low correlation indicates that SBERT and human similarity perception might rely on different aspects of sentence similarity.

Preferred Scores. The average similarity score computed for the SimilEx dataset is 2.40 (± 0.98), which suggests that the paired sentences are often perceived as different by their annotators. As proof, consider Figure 1, which illustrates the percentage distribution of mean scores of SimilEx pairs, computed by averaging the scores assigned by individual participants. Most pairs (76.86%) received scores <3, the midpoint of the scale, while only 7.05% of sentence pairs obtained a mean score ≥ 4. Consistent with these findings, scores 4 and 5, indicating similarity, account for only 23.46% of the individual scores assigned during the campaign by participants. In contrast, scores 1 and 2, indicating dissimilarity, are much more prevalent (57.59%). The neutral score of 3 is also relatively common (16.76%), suggesting that in many cases subjects could not decisively determine the similarity of the paired sentences.

Inter-annotator Agreement. To explore the consistency of these perceptions, we examine the inter-annotator agreement (IAA) on the similarity scores using Krippendorff's α coefficient, a metric suitable when the items have a different number of annotations.

The global IAA, computed considering all pairs and annotators, is 0.352. A *fair* [25] agreement is not surprising due to the inherent subjectivity of the task, yet it still indicates a tendency for annotators to converge on many items. To explore this further, we grouped sentence pairs based on the number of annotators who assigned them the same score. The resulting groups have quite different sizes: more than half of the pairs (around 56%) have 3 or fewer annotators in agreement, while 5 or more annotators (up to 9) gave identical values in 20.08% of pairs. Figure 2 displays the distribution of similarity scores within these groups. Notably, when few annotators agree on a pair, the scores are evenly distributed across the five labels, indicating that disagreement can occur for pairs seen as both similar and different. In contrast, when more annotators agree, the most commonly assigned score is 1, indicating that annotators converge more frequently on dissimilarity judgments. This is supported by the negative Pearson correlation between the number of agreeing annotators and the average similarity score of the pair ($r = -0.344$, $p < 0.001$).

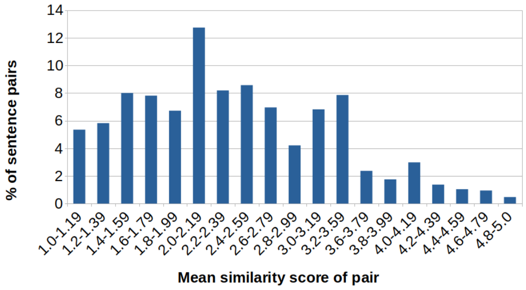

Figure 1: Percentage distribution of mean similarity scores of SimiEx pairs.

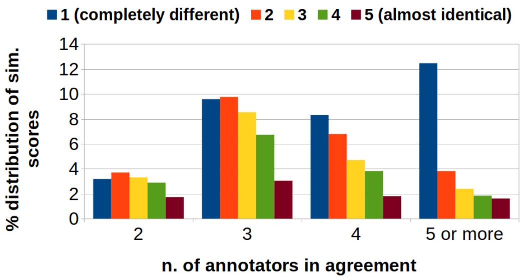

Figure 2: Percentage distribution of similarity scores with respect to the number of annotators in agreement on the pair.

Agreement, Style and Similarity. We explored the relationship between style and similarity judgments by comparing scores and stylistic traits of sentences. As a general remark, we found that style minimally affects pairs' similarity: the Pearson correlation between the similarity scores and the distribution of stylistic properties is either non-significant ($p > 0.05$) or extremely low ($r < 0.1$). However, a more in-depth analysis of specific stylistic properties revealed a nuanced relationship between style and the consistency of human judgments. For example, contrary to our expectations, sentence length, a raw yet informative feature reflecting stylistic variation, did not impact the similarity scores assigned by annotators. In fact, when we computed the correlation between the length difference of paired sentences and the variance between similarity judgments, we observed a lack of correlation (0.05). To further investigate, we grouped pairs based on the difference between the length of their sentences, and specifically, based on whether their length difference was above or below the average value of 17 tokens. We noticed that also from this perspective of analysis sentence length did not affect the IAA of the scores either, as $\alpha = 0.265$ for both groups. However, when focusing on different stylistic traits more closely related to sentence structure, we observed a substantial relationship with higher annotator agreement. For instance, the IAA is moderate (0.49) for pairs where neither sentence contains a subordinate clause, but drops to fair (0.25) when both sentences contain at least one subordinate. Similarly, the IAA is higher (0.37) when the syntactic tree depth difference between paired sentences is below the average value of 1.98, compared to 0.29 when the difference is greater. These results are extremely interesting as they indicate that while stylistic traits may not directly influence the semantic similarity between sentences, some of them play a role in the convergence of human judgments.

4. Human Similarity Explanation

In this section, we focus on the analysis of the subset of 907 sentence pairs of SimilEx annotated by the two students with both human similarity judgments and natural language explanations for the assigned scores.

Comparison with Prolific annotators. The comparison between the similarity judgments of the graduate students and Prolific annotators reveals a strong alignment between the two groups. The Pearson correlation between the average similarity score of the Prolific annotators and the average score between the two graduate students is significantly high and positive ($r = 0.779$, $p < 0.001$). This high correlation is also observed when computed separately for each of the two students, indicating that their perceptions of similarity closely match the judgements obtained from the crowdsourcing campaign. Additionally, the IAA between the two students suggests alignment between the students since $\alpha = 0.49$, higher than that reported among the Prolific annotators.

Linguistic Style of Explanations. We explored the style of explanation relying on the linguistic profiling method described in Section 2.1. We noted that the explanations written by the two students exhibit partial similarity as can be seen by inspecting the results of the stylistic analysis distributed as supplementary materials (see Appendix A). For example, they both tend to write quite short sentences, i.e. on average 6.35 (±3.93) and 7.67 (±5.12) token-long, and characterized by a nominal style. This is evidenced by the low percentage distribution of verbal roots (i.e. sentences with a verb as the syntactic root), computed over the total number of roots represented by other morpho-syntactic categories (i.e. 58.21% (±49.35) and 61.43% (±48.70)). This percentage is notably low when compared to the distribution in the ISDT [26], the largest Italian Treebank, where the distribution is 85.73%.

Content of Explanations. The content analysis of the explanations reveals that both students share some arguments when justifying the similarity scores for SimilEx sentence pairs. Specifically, the average cosine similarity between their explanations, computed using SBERT, is 0.46, indicating a moderate level of similarity.

Given that a qualitative analysis reveals several recurring arguments and templates in the explanations, such as *Entrambe descrivono* ('Both describe'), *In entrambe le frasi si parla di un argomento militare* ('In both sentences a military topic is mentioned'), we further explored the possibility of identifying homogenous content among them. To this end, we clustered the 907 explanations of each student (1,814 in total) based on their SBERT vectors. We initially configured the clustering algorithm to partition the data into 10 clusters[5]. However, only 4 of these clusters were found to be semantically homogeneous. Specifically, these homogeneous clusters contain explanations where either Student 1 or 2: *i)* writes that the evaluated sentences contain positive or negative emotions such as love or anger, *ii)* uses the phrase *Pressochè identiche* ('Almost identical'), *iii)* uses the phrase *Completamente diverse* ('Completely different'), and *iv)* notes that the evaluated sentences refer to a military topic. Since the explanations in the remaining 6 clusters were not semantically homogeneous, we reconfigured the clustering algorithm to partition the data into 5 clusters. This time, we included only the explanations that had not been previously clustered, representing 72.76% of all SimilEx explanations. However, we were still unable to isolate explanations with similar content. This suggests that the two students often focused on different aspects when evaluating sentence similarity. As proof, consider the examples reported in Table 2, where stu-

[5] We employed agglomerative clustering using Euclidean distance and Ward variance minimization as the clustering method.

(1)	**Sentence 1**	*"Vedeva lo scintillio degli occhi, tremulo e avvampante, e il riso di felicità e di eccitamento che senza volere le increspava le labbra; vedeva la grazia misurata, la sicurezza e la levità dei movimenti."*
	Sentence 2	*"Era così bella, che non solo non appariva in lei ombra di civetteria, ma pareva al contrario che le rimordesse il forte ed immancabile effetto di una grazia trionfatrice, che avrebbe voluto temperare, se le fosse stato possibile."*
	Explanations	**S1:** Completamente diverse. (1)
	(Sim. scores)	**S2:** Parlano di donne che sono molto graziose. (4)
(2)	**Sentence 1**	*"Ma che volete farci: questa è la vocazione dell'autore, ormai malato della propria imperfezione, e il suo talento è fatto apposta per rappresentare la povertà della nostra vita, scovando la gente in buchi sperduti, in angoletti remoti dell'impero!"*
	Sentence 2	*"Perché mettere in mostra la povertà della nostra vita e la nostra triste imperfezione, andando a scovare gli uomini in buchi sperduti, in angoletti remoti dell'impero?"*
	Explanations	**S1:** Completamente diverse anche se esprimono lo stesso concetto. (1)
	(Sim. scores)	**S2:** Stessa frase impostata diversamente a livello sintattico. (4)
(3)	**Sentence 1**	*"L'agente di polizia che l'accompagnava, discese e scosse il braccio intormentito; poi si tolse il berretto e si fece il segno della croce."*
	Sentence 2	*"Nell'osteria entrò un agente di polizia."*
	Explanations	**S1:** In entrambe le frasi si parla di un agente della polizia. (2)
	(Sim. scores)	**S2:** Il soggetto è un agente di polizia. (2)
(4)	**Sentence 1**	*"Napoleone si volse ad Alessandro, come per dire che quanto ora faceva era fatto per l'augusto e caro alleato."*
	Sentence 2	*"Tutti gli alleati di Napoleone gli divennero nemici."*
	Explanations	**S1:** In entrambe le frasi si parla di Napoleone e dei suoi alleati. (2)
	(Sim. scores)	**S2:** Parlano degli alleati di Napoleone. (3)
(5)	**Sentence 1**	*"Ma l'amore con un marito inquinato dalla gelosia e da ogni sorta di difetti non era più per lei."*
	Sentence 2	*"Era forse, semplicemente, un sentimento di gelosia: egli era talmente avvezzo all'amore di lei, che non poteva ammettere che ella potesse amarne un altro."*
	Explanations	**S1:** Nel primo caso il focus della frase è la moglie, nella seconda lo è il marito. (2)
	(Sim. scores)	**S2:** Parlano di uomini gelosi. (2)
(6)	**Sentence 1**	*"Tonfi, spruzzi, strida, ingiurie, lazzi, risate, un allegro pandemonio."*
	Sentence 2	*"E fino a quel momento, chiasso, baccano, sghignazzi, ingiurie, rumore di catene, acido carbonico e fuliggine, teste rase, facce marchiate, vestiti a brandelli, tutto fatto oggetto di ludibrio e di infamia... sì, grande è la vitalità dell'uomo!"*
	Explanations	**S1:** Entrambe le frasi descrivono vitalità. (4)
	(Sim. scores)	**S2:** Descrivono degli scenari di caos, disordine; sintassi frasi simile. (3)

Table 2

Sentence pairs with similarity scores and explanations (translations in App. D). Examples 1-2 illustrate divergent explanations and scores; 3-6 show identical or aligned scores, with explanations mentioning similar (3-4) or different (5-6) aspects.

dents focused on diverse aspects of the paired sentences while they assigned either similar (see #5 and #6) or different (see #1 and #2) similarity scores. While this may result in underspecification and inconsistency in the collected explanations, it confirms the inherent subjectivity and expressivity involved in providing free-text natural language explanations for a highly subjective task such as evaluating semantic sentence similarity [10].

The content analyses above were enriched with an in-depth investigation into whether there is a correlation between the SBERT cosine similarity of the explanations of each student and their similarity judgments. The Pearson correlation between SBERT scores and the absolute difference in the students' similarity judgments reveals a moderate negative relationship ($r = -0.459$, $p < 0.001$). This indicates that the more semantically similar the explanations are, the smaller the difference in the students' similarity judgments. Notably, students' explanations tend to be more similar when the similarity scores assigned by both of them are lower (i.e. 1 or 2), as in example #3 of Table 2.

5. Conclusion and Future Work

This paper presented SimilEx, the first Italian dataset on sentence similarity enriched with human judgments and free-form explanations. The analyses of the collected judgments confirmed that the perception of sentence similarity is inherently subjective, as evidenced by the fair agreement between the scores. Notably, annotators tend to agree less on similar sentence pairs, showing greater convergence when sentences are markedly different. The style of the paired sentences appears to influence this convergence: while most linguistic traits may not directly impact the similarity score, some of them affect the homogeneity of judgments assigned by different annotators. These features mostly concern properties of sentence structure rather than raw sentence features such as lenght, which does not play a role in homogeneity. Regarding explanations, we found a correlation between the similarity of the content of the explanations and the similarity scores assigned, indicating that annotators tend to write more similar explanations, using a similar writing style, when their scores align.

The findings from this study open several prospects. Expanding SimilEx to include sentences from different textual genera could provide further insights into the factors affecting similarity judgments. Additionally, incorporating more annotators with varying linguistic backgrounds could foster a better understanding of the subjectivity in similarity perception. Lastly, our dataset could help develop automated tools to evaluate the explainability of LLMs. By leveraging SimilEx, researchers can create models that predict similarity scores and generate explanations, enhancing the interpretability of LLMs.

6. Acknowledgments

This paper is supported by the PRIN 2022 PNRR Project EKEEL - Empowering Knowledge Extraction to Empower Learners (P20227PEPK) funded by the European Union – NextGenerationEU and the project CHANGES - Cultural Heritage Innovation for Next-Gen Sustainable Society (PE00000020) project under the NRRP MUR program funded by the NextGenerationEU.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[3] G. Gemini Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[4] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1906–1919.

[5] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, et al., Factuality challenges in the era of large language models, arXiv preprint arXiv:2310.05189 (2023).

[6] Y. Belinkov, J. Glass, Analysis methods in neural language processing: A survey, Transactions of the Association for Computational Linguistics 7 (2019) 49–72.

[7] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[9] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–38.

[10] S. Wiegreffe, A. Marasovic, Teach me to explain: A review of datasets for explainable natural language processing, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[11] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).

[12] S. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.

[13] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).

[14] A. Brassard, B. Heinzerling, P. Kavumba, K. Inui, COPA-SSE: Semi-structured explanations for commonsense reasoning, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022, pp. 3994–4000.

[15] A. Zaninello, S. Brenna, B. Magnini, Textual entailment with natural language explanations: The italian e-RTE-3 Dataset, in: F. Boschetti, alii (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), November 30 - December 2nd, Venice (Italy), 2023.

[16] J. Wang, Y. Dong, Measurement of text similarity:

a survey, Information 11 (2020) 421.

[17] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, * sem 2013 shared task: Semantic textual similarity, in: Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity, 2013, pp. 32–43.

[18] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, D. Jurgens (Eds.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14.

[19] Y. Wang, S. Tao, N. Xie, H. Yang, T. Baldwin, K. Verspoor, Collective Human Opinions in Semantic Textual Similarity, Transactions of the Association for Computational Linguistics 11 (2023) 997–1013.

[20] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[22] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), ELRA, 2020, pp. 7147–7153.

[23] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. doi:10.1162/coli_a_00402.

[24] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, *SEM 2013 shared task: Semantic textual similarity, in: M. Diab, T. Baldwin, M. Baroni (Eds.), Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 32–43. URL: https://aclanthology.org/S13-1004.

[25] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

[26] C. Bosco, S. Montemagni, M. Simi, Converting italian treebanks: Towards an italian stanford dependency treebank, in: Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse, 2013.

Appendix

A. Supplementary materials

The complete SimilEx dataset is freely available at http://www.italianlp.it/resources/ along with the results of the stylistic analysis of both paired sentences and the natural language explanations provided by the two students.

Specifically, on the dedicated page, you can find the following materials:

SimilEx dataset. The dataset is organized in columns, each reporting the following information:

- Pair_ID: the unique identifier of the paired sentences;
- Sentence_1 and Sentence_2: the text of each of the two paired sentences;
- A1-A7: the similarity judgments of the Prolific annotators;
- Stud_1: the similarity judgment assigned by the first student;
- Explanation_Stud1: the natural language explanation provided by Stud_1;
- Stud_2: the similarity judgment assigned by the second student;
- Explanation_Stud2: the natural language explanation provided by Stud_2.

Linguistic profiling of the paired sentences. The results of the stylistic analysis of each of the paired sentences included in SimilEx are contained in the "Sentence_profiling" sheet, reporting for each column the following information:

- Pair_ID: the unique identifier of the paired sentences in the SimilEx dataset;
- Sent_in_pair: the unique identifier of each individual sentence in the pair;
- all other columns report the value of the distribution of the complete set of linguistic characteristics derived with Profiling-UD by each individual sentence.

Linguistic profiling of the explanations. The results of the stylistic analysis of each explanation provided by the two students are contained in the "Explanations_profiling" sheet, reporting for each column the following information:

- PairID_of_explanied_pair: the unique identifier of each individual sentence in the pairs of the SimilEx dataset;

- Explanation_of_student: the identifier of the student;
- all other columns report the value of the distribution of the complete set of linguistic characteristics derived with Profiling-UD by each individual explanation.

B. Linguistic Features

The set of linguistic features derived by Profiling–UD are extracted from different levels of linguistic annotation and capture a wide number of linguistic phenomena and can be grouped as follows:

- **Raw text**
  - Number of tokens in sentence;
  - Average characters per token.
- **Morphosyntactic information**
  - Distibution of UD POS;
  - Lexical density.
- **Inflectional morphology**
  - Distribution of lexical verbs and auxiliaries for inflectional categories (tense, mood, person, number).
- **Verbal Predicate Structure**
  - Distribution of verbal heads and verbal roots;
  - Average verb arity and distribution of verbs by arity.
- **Global and Local Parsed Tree Structures**
  - Average depth of the whole syntactic trees;
  - Average length of dependency links and of the longest link;
  - Average length of prepositional chains and distribution by depth;
  Average clause length.
- **Relative order of elements**
  - Distribution of subjects and objects in post- and pre-verbal position.
- **Syntactic Relations**
  Distribution of dependency relations.
- **Use of Subordination**
  - Distribution of subordinate and principal clauses;
  - Average length of subordination chains and distribution by depth;
  - Distribution of subordinates in post- and pre-principal clause position.

C. Annotation Instructions

C.1. Original Instructions in Italian

Stai per svolgere un questionario nel quale ti verrà chiesto di valutare se due frasi sono fra di loro simili o diverse. Per farlo, ti mostreremo delle coppie di frasi estratte da romanzi e ti chiederemo di assegnare ad ogni coppia un punteggio compreso fra 1 e 5.
Usa 1 per dire che le due frasi sono fra loro completamente diverse; Usa 5 per dire che sono pressoché uguali. Gli altri punteggi ti serviranno per valutare i casi intermedi.

Due frasi possono dirsi uguali o diverse sulla base di diversi elementi. Ecco alcuni esempi per aiutarti nella valutazione.

Coppie di frasi diverse (punteggio 1).

Esempio 1:
a) Io desidererei tanto non sentire così intensamente e non prendermi tanto a cuore tutto quello che succede.
b) Sì, non sono in me, sono tutta nell'aspettativa e vedo tutto un po' troppo facile.

Esempio 2:
a) Anche il vecchio principe t'è affezionato.
b) - Non mi sembra di averveli chiesti, - scattò il principe irritatissimo.

Esempio 3:
a) Il the veramente era del color della birra, ma io ne bevvi un bicchiere.
b) Ma non passò neanche un minuto, che la birra gli diede alla testa e per la schiena gli corse un leggero e perfin piacevole brivido.

Fai particolare attenzione agli esempi 2 e 3: anche se le frasi hanno delle parole in comune (come 'principe' e 'birra' negli esempi) non è detto che siano uguali!

Coppie di frasi molto simili (punteggio 5).

Esempio 1:
a) Signori della giuria, la psicologia è a doppio taglio e anche noi siamo in grado di comprenderla.
b) Vedete allora, signori della giuria, dal momento che la psicologia è un'arma a doppio taglio, permettetemi di occuparmi del secondo taglio e vediamo che cosa viene fuori.

Esempio 2:
a) "Un rettile divorerà l'altro", aveva detto il giorno prima Ivan, parlando con rabbia del padre e del fratello.
b) "Un rettile divorerà l'altro, quella è la fine che faranno!".

Esempio 3:
a) Ma una volta deciso, continuò con la sua voce stridula, senza timori, senza esitazioni e sottolineando alcune parole.
b) Parlava rapido, senza fermarsi un momento, senza la minima esitazione, quasi rimproverasse a sè stesso di aver tanto indugiato a mettere Marianna a parte di tutti i suoi segreti, quasi scusandosi presso di lei.

Gli esempi 1 e 2 riportano frasi che non solo contengono molte parole in comune ma sono simili anche per quanto riguarda la scena descritta. Nel terzo esempio, entrambe le frasi descrivono una persona intenta a parlare in modo svelto e deciso. Possiamo dire che in questi esempi l'alta similarità fra le frasi è data dal fatto che, ad eccezione di alcuni dettagli, esse descrivono scene o immagini molto simili, anche se si svolgono in contesti diverse.

C.2. Instructions Translations into English

You are about to take a questionnaire in which you will be asked to assess whether two sentences are similar or different to each other. To do this, we will show you pairs of sentences extracted from novels and ask you to give each pair a score between 1 and 5.
Use 1 to say that the two sentences are completely different from each other; Use 5 to say that they are almost the same. The other scores will be used to evaluate the intermediate cases.

Two sentences can be equal or different based on several elements. Here are some examples to help you in your evaluation.

Pairs of different sentences (score 1)

Examples: Please refer to the above section to see the original examples in Italian.

Pay particular attention to examples 2 and 3: although the sentences have words in common (like 'prince' and 'beer' in the examples) they are not necessarily the same!

Pairs of very similar sentences (score 5)

Examples: Please refer to the above section to see the original examples in Italian.

Examples 1 and 2 show sentences that not only contain many words in common but are also similar in terms of the scene described. In the third example, both sentences describe a person speaking quickly and decisively. We can say that the high similarity between the sentences in these examples is due to the fact that, except for a few details, they describe very similar scenes or images, even though they take place in different contexts.

D. Translations of Explanations

English translations of the similarity explanations originally written in Italian by the two students and reported in Table 1.

- **Example (1)**
  S1: Completely different.
  S2: They talk about women who are very pretty.
- **Example (2)**
  S1: Completely different although they express the same concept.
  S2: Same sentences with different syntactic structures.
- **Example (3)**
  S1: In both sentences, a police officer is mentioned.
  S2: The subject is a police officer.
- **Example (4)**
  S1: In both sentences, Napoleon and his allies are mentioned.
  S2: They speak of Napoleon's allies.
- **Example (5)**
  S1: In the first case the focus of the sentence is the wife, in the second it is the husband.
  S2: They talk about jealous men.
- **Example (6)**
  S1: Both sentences describe vitality.
  S2: They describe scenarios of chaos, disorder; similar sentence syntax.