Unraveling Text Coherence from the Human Perspective: a Novel Dataset for Italian

Federica Papa[1], Luca Dini[1,2], Dominique Brunato[2] and Felice Dell'Orletta[2]

[1] *University of Pisa*
[2] *Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa*

Abstract
This paper presents a novel resource designed to study text coherence in the Italian language. The dataset aims to address existing deficiencies in coherence assessment by focusing on human perception of coherence. Recently, it has been integrated into the DiSCoTex benchmark, part of EVALITA 2023 [1], the 8th evaluation campaign for NLP and speech tools in Italian. Our resource aims to provide a comprehensive understanding of coherence, highlighting the influence of both genre and text perturbations on perceived coherence.

Keywords
text coherence, human perception, Italian dataset, text perturbations

1. Introduction and Motivation

Coherence plays a central role in maintaining the overall unity of a text and is influenced by both linguistic and extra-linguistic factors. From the linguistic point of view, it primarily relies on cohesion, which encompasses various linguistic devices used in natural languages to establish connections within a text, such as anaphoric and cataphoric relationships, discourse markers, and elliptical constructions [2]. While cohesion mainly ensures local coherence between adjacent or nearby sentences, to be fully coherent a text needs to achieve a global coherence, a property that pertains to the connection of concepts and relationships that underlie the surface text ensuring a logical flow of ideas around an overall intent [3]. This aspect of coherence adds a subjective component, as it also depends on the reader or listener's familiarity with the text, language proficiency, and level of interest and attention.

Modelling coherence in natural language is essential for a wide range of downstream applications. One such application is automatic essay scoring in language learning settings, where coherence assessment can provide valuable writing feedback by identifying poorly organized paragraphs and abrupt topic transitions [4, 5]. In clinical contexts, coherence modeling is relevant for automatic language assessment, as speech irregularities indicative of a lack of coherence can serve as markers for mental disorders like schizophrenia [6, 7]. Furthermore, coherence has been adopted as an intrinsic evaluation metric for assessing the quality of texts generated by Natural Language Generation (NLG) systems [8]. Additionally, coherence modeling is gaining importance in research on the interpretability of modern deep neural networks [9, 10, 11]. Indeed, while existing work has mainly focused on probing sentence-level properties, understanding how these models encode discourse and pragmatic phenomena remains a crucial aspect.

In light of this interest, various attempts have been made to approach coherence assessment in the Natural Language Processing (NLP) community, especially in the 'pre-deep learning' era. With this respect, early computational models of discourse coherence were primarily built upon two linguistic theories: centering theory [12] and rhetorical structure theory [13] . Studies aligned with centering theory, such as [14], focused on analyzing the distribution of entity transitions over sentences as a means to predict text coherence. On the other hand, works inspired by rhetorical structure theory, such as [15], employed discourse parsers to generate discourse relations over sentences. With the advent of neural models, researchers have also explored their application in coherence assessment, see e.g. Lin et al. [15] and Nguyen and Joty [16].

The importance of building challenging datasets for coherence evaluation cannot be overstated. With this respect, independently from the underlying theories, models of discourse coherence are typically tested on tasks such as reordering, which aim to discern an original text from a corrupted one artificially created by shuffling the order of its sentences, or tasks that require systems to detect whether a document contains an intruder sentence from another document [9] or to classify whether a target sentence is contiguous or not with a given passage [17]. However, these approaches have come under criticism because they neglect key aspects of coherence, as

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

f.papa4@studenti.unipi.it (F. Papa); luca.dini16@gmail.com (L. Dini); {dominique.brunato}@ilc.cnr.it (D. Brunato); {felice.dellorletta}@ilc.cnr.it (F. Dell'Orletta)

noted by Lai and Tetreault [4] and Beyer et al. [18] among others. They fail to identify the qualities that make the shuffled text incoherent, do not pinpoint the linguistic devices responsible, and overlook the subjective component underlying coherence. Additionally, most existing benchmarks are limited to the English language.

Our contribution In this paper we seek to address some of the existing deficiencies in coherence assessment by presenting a novel resource tailored for the Italian language, designed specifically to study text coherence from the perspective of human perception. The dataset, which to our knowledge is the first for Italian, has been recently used as part of a larger benchmark released for DiSCoTex, one of the shared-tasks presented at the 8th evaluation campaign of NLP and speech tools for the Italian language (EVALITA 2023) [1]. The results of first analyses on the resource shed light on the influence of both genre and text perturbations on perceived coherence[1].

2. Dataset Construction

The construction of our dataset was guided by two distinct criteria: on the one hand, we intended to explore the effect of textual genre on the human perception of coherence; on the other hand, we wanted to assess whether and to what extent humans are sensitive to different strategies introduced to artificially modify an original text.

As a starting point we selected texts from two distinct sources: the Italian Wikipedia and the Italian speech transcripts section of the Multilingual TEDx corpus (mTEDx). The choice of these sources was meant to obtain a balanced corpus that was representative of two different language varieties: the former is a 'standard' written variety, and the latter a 'hybrid' variety combining diverse genres (e.g., university lectures, newspaper articles, conference presentations, and TV science programs) as well as different semiotic modes, such as written, spoken, audio, and video [19].

Following the approach by Brunato et al. [17], for each text we then proceed to extract passages consisting of four consecutive sentences, considering them as our unit of analysis for modeling the coherence annotation task. As for Wikipedia, we relied on the existing segmentation into paragraph and extract four-sentence passages. For the TEDx corpus, as these texts lack such an internal structure, we split all the transcripts into passages of four sentences.

After creating all the possible passages, we randomly selected 1,064 of them while maintaining a proportional representation from both sources. Half of the extracted passages were left unchanged, while the other half underwent a perturbation. More specifically, we devised two distinct perturbation strategies:

- *swap*: it involves swapping the position of two random sentences in the text passage.
- *substitution* (sub): it consists of replacing one of the four sentences with another sentence, corresponding to the 10th sentence following the passage in the same document.

Table 1 contains an example from the corpus for each perturbation type

2.1. Collecting human ratings

Before starting the annotation process with humans, we added ten fillers to the dataset, consisting of four-sentence passages deliberately chosen to be either highly coherent or highly incoherent. These additional passages served as a control mechanism to check the reliability and accuracy of each annotator in assigning coherence scores to the actual texts in the dataset: if an annotator assigns an out-of-scale coherence value to these texts, it suggests that they might not have conducted the annotation process adequately.

The annotation process has been executed via *crowd-sourcing*. We first used the *Questbase*[2] platform to create questionnaires formulating the text scoring process in the form of questions. Then, we distributed the questionnaires using the crowdsourcing platform *Prolific*[3], choosing to recruit only Italian native speakers without language disorders as annotators. Considering that subjective component underlying coherence that makes this concept gradual rather than categorical, people were asked to rate each texts on a 5-point Likert scale, where 1 represents the minimum value of perceived coherence and 5 the maximum[4]. A pilot experiment tested the suitability of the questionnaire from different points of view (i.e. the clearness of instructions) and allowed to estimate the time needed to complete it. After collecting all the responses to the questionnaires, we kept only the most reliable annotations by filtering out the annotators who had failed the attention checks. Specifically, we excluded those annotators who rated at least four control texts incorrectly, i.e. assigning a value from 1 to 3 to highly coherent filler passages or a value from 3 to 5 to very incoherent filler passages. As a result, we retained an average of 10 annotations per passage for a total of 10,567 annotations for the whole dataset.

[1] The dataset will be made publicly available for research purposes at the following link: http://www.italianlp.it/resources/

[2] https://questbase.com/

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

[4] Appendix A contains the instructions given to the annotators when opening the questionnaire on Prolific.

Table 1
Example of perturbations: in the first example, which is a passage from Wikipedia, sentence 2 and sentence 3 have been swapped. In the second one, from the TEDx corpus, the 4th sentence have been substituted with the subsequent 10th sentence.

Text passage	Perturbation
1. Cliff Burton possedeva uno stile impeccabile ed era capace di produrre giri di basso potenti ma allo stesso tempo raffinati. **2.** Specialmente durante gli assoli era solito pizzicare due o tre corde nello stesso momento e lanciarsi in un complesso uso di distorsioni, tapping, bending e applicazioni del pedale wah wah. **3.** Il suo stile era molto vario per i canoni di un bassista heavy metal: Burton non suonò mai il basso "come un chitarrista" e mai utilizzò plettri, prediligendo il contatto diretto con le corde, pizzicate a mani nude. **4.** Anche per questo, diversamente da altri bassisti heavy metal che utilizzavano bassi a cinque o sei corde, Burton suonava solo bassi a quattro corde, che considerava più adatti al suo stile.	Swap 2-3
1. È stato teorizzato che le prime stelle dell'universo, le cosiddette stelle di Popolazione III, fossero molto più massicce delle stelle attualmente esistenti. **2.** Si è postulata l'esistenza di questa prima generazione di stelle per spiegare l'esistenza di elementi chimici diversi dall'idrogeno e dall'elio nelle stelle più vecchie conosciute. **3.** Sebbene fossero più grandi e luminose di tutte le supergiganti note oggi, la loro struttura doveva essere molto differente, con perdite di massa molto più contenute. **4.** Nella maggior parte dei casi la variabilità è dovuta a pulsazioni della superficie stellare.	Sub 4

3. Analysis of perceived coherence

To delve deeper on the factors influencing human perception of text coherence, we conducted two types of analyses that examine the relationship between perceived coherence and text structure from distinct perspectives. The first one focuses on the effect of the different perturbation strategies artificially introduced to disrupt the internal coherence of rated passages; the second one takes into account solely the subset of original, i.e. unperturbed, texts with the aim of exploring the effect of several linguistic features extracted from each passage on the mean coherence judgments.

3.1. Impact of text perturbations

After gathering all annotations, we studied their homogeneity by calculating for each passage the mean value and standard deviation of the coherence scores assigned to it[5]. These statistics were computed for the whole dataset as well as for passages grouped according to the text source from which they derived (TED or Wikipedia) and to the perturbations eventually applied. The purpose was to observe how coherence ratings vary among the different groups and understand the effects of the different artificial perturbations applied to the text. These results are shown in Figure 1.

Observing the trend of the distribution of the mean coherence ratings for each group, it was possible to see that the group containing all the original texts was considered as more coherent than the ones with the perturbed texts. However, in all considered groups, texts extracted from Wikipedia were rated as more coherent than those extracted from TEDx, even when artificially perturbed. This suggests that Wikipedia documents tend to exhibit a more standardized structure, with internal coherence remaining relatively stable even when subjected to minor alterations, such as changes in sentence order or the inclusion of an intruder sentence from the same document.

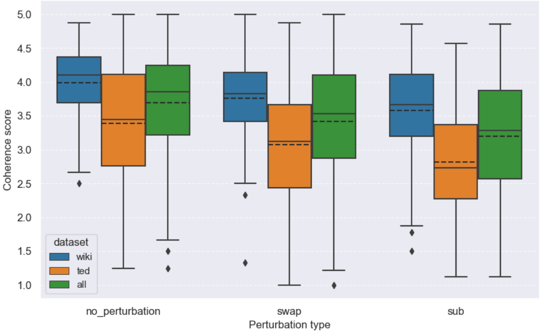

Figure 1: Box plot of human judgments collected for the dataset. For each subset (*Wiki*, *TED*, and all) and the possible perturbations (*Sub*, *Swap*). The dashed line corresponds to the mean coherence score.

In all groups the standard deviation has low values, suggesting a high degree of homogeneity in the ratings, especially for Wiki passages.

A more in-depth investigation was conducted to assess the potential impact of certain factors of each perturbation type, such as the **distance** between swapped sentences and the **position** of the replaced sentence, on the distribution of mean coherence scores. Regarding the swap perturbation, we assigned a label to each perturbed text, indicating the distance between the swapped sentences. Dist_0 was given to passages where the swapped sentences were adjacent, Dist_1 to those with one sentence between the swapped sentences, and Dist_2 to

[5] Inter-annotator agreement measured by Krippendorff's alpha is .32 for the whole corpus.

those with two sentences between the swapped sentences. Similarly, we assigned a label to the passage that underwent the substitution perturbation. Pos_1 was given to passage where the first sentence was replaced, Pos_2 to those where the second sentence was replaced, Pos_3 when the third sentence was replaced, and Pos_4 when the fourth sentence was replaced.

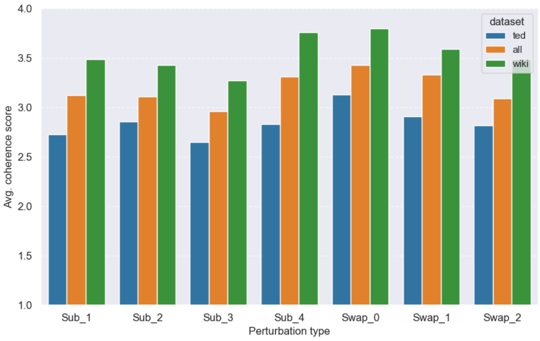

Figure 2: Mean coherence judgments attributed to perturbed passages grouped by distance (for *swap* perturbation) and position (for *substitution* perturbation).

As we can see in Figure 2, in texts perturbed with the swap perturbation, the perceived coherence is higher in those cases where the sentences are adjacent to each other, while decreases as the distance between the exchanged sentences increases. In texts altered via substitution, the perceived coherence increases especially in those cases where the sentence substitution occurred at the ends (thus in the first and last positions), while it generally decreases in cases where substitution involves the middle positions.

3.2. Impact of linguistic structure

Although the previous analysis revealed that perturbed texts were rated on average as less coherent that original ones, we also observed that such a perception is influenced by textual genre. Considering the subjective nature of coherence, we hypothesize that even well-formed texts may receive different coherence annotations. We thus carried out a final analysis focused solely on the subset of original texts, with the aim of investigating the relationship between the linguistic profile of these texts and the perceived coherence.

To automatically extract linguistic information from human rated passages, we leveraged *Profiling-UD* [20], a tool for carrying out linguistic profiling investigations in multiple languages based on the Universal Dependency framework. Using *Profiling-UD*, we extracted more than 130 features for each passage, which capture lexical, morpho-syntactic and syntactic properties of text. An overview of these features is shown in Table 2. These features were shown to be relevant for modeling aspects characterizing the interaction between a reader and a text, such as the human perception of sentence complexity [20] and of writing quality [21]. As both complexity and quality are properties connected to coherence, we expect that these features will provide valuable insights for our coherence analysis as well.

We then sought correlations between the human perception of coherence and these linguistic features by calculating the Spearman correlation coefficient between the average coherence score attributed to each passage and the average value of each linguistic feature extracted from it. Results are shown in Table 3. Considering only features with *p-value* below the threshold of 0.05, we observed that the perception of coherence in original texts positively correlates above all with features closely related to *length.* In fact, the highest correlation (Spearman's r = 0.32) was obtained with the maximum depth of syntactic tree, followed by *tokens_per_sent* and *n_tokens*, which captures respectively the average sentence length and of passage length in number of tokens. Also *lexical richness*, measured by the average value of Type/Token Ratio (*ttr_form_100*) turned out to be among the first top-five correlated features. These findings suggest that longer sentences could contain more information and thus lead to a more complete text, that also makes it more coherent. An interesting result was that the use of pronouns negatively correlated with the perception of coherence (Spearman's r = -0.26). This unexpected observation can be attributed to the potential ambiguity of pronouns, deriving from the fact that the evaluated passages were extracted from larger texts and might lack the necessary context to accurately link the pronoun to its intended referent.

Focusing specifically on original passages derived from Wikipedia, we observed that the presence of features describing proper syntactic phenomena closely related to the sentence length, such as the average length of dependency links and of the maximum link (*avg_links_len*, *max_links_len*), along with the presence of nouns modified by prepositional phrases (*prep_chain*), contributed to increased coherence perception by annotators. These linguistic features that are typically related to syntactic complexity may suggest that these texts are also more informative, resulting in enhanced coherence perception. Furthermore, it could be seen that in texts taken from Wikipedia the judgement of coherence was positively influenced by a paratactic structure of the text (*dep_conj*). Finally, in the TED original texts it could be observed that the correlation was positive in the case of the distribution of subordinate propositions (*subord_dist*) and negative in the distribution of main ones (*princ_dist*).

Table 2
Overview of linguistic features used by Profiling-UD.

Annotation Level	Linguistic Feature Description	Label
Raw Text	Sentence length (tokens), word length (characters) Words and lemmas type/token ratio	n_tokens, char_per_tok ttr_form, ttr_lemma
POS Tagging	Distribution of UD and language-specific POS tags Lexical density Inflectional morphology of auxiliaries (mood, tense)	upos_dist_*, xpos_dist_* lexical_density aux_mood_*, aux_tense_*
Dependency Parsing	Syntactic tree depth Average and maximum length of dependency links Number and average length of prepositional chains Relative ordering of main elements Distribution of dependency relations Distribution of verbal heads Distribution of principal and subordinate clauses Average length of subordination chains Relative ordering of subordinate clauses	parse_depth avg_links_len, max_links_len n_prep_chains, prep_chain_len subj_pre, subj_post, obj_pre, obj_post dep_dist_* vb_head_per_sent princ_prop_dist, sub_prop_dist sub_chain_len sub_post, sub_pre

Table 3
Extract of linguistic features correlating with mean coherence judgments attributed to all original passages (Unp_ALL), original Wiki-extracted passages (Unp_WIKI) and original TED-extracted passages (Unp_TED). Significant correlations (*p-value* < 0.05) are denoted with a star.

LingFeats	Unp_ALL	Unp_WIKI	Unp_TED
avg_max_depth	0.32*	0.09	0.33*
tok_per_sent	0.3*	0.11	0.21*
n_tok	0.28*	0.14	0.16*
upos_ADP	0.26*	-0.03	0.23*
ttr_form_100	0.26*	0.16*	0.14
upos_ADV	-0.25*	-0.08	-0.13
upos_PUNCT	-0.26*	-0.04	-0.28*
upos_PRON	-0.26*	-0.1	-0.06
max_links_length	0.19	0.24*	0.04
verb_tense-Past	0.24*	0.21*	0.01
prep_chain	0.26*	0.19*	0.18
avg_links_len	0.05	0.19*	-0.08
aux_mood_Ind	0.034	0.18*	0.02
aux_form_Fin	-0.05	0.15*	-0.19*
dep_conj	-0.14	0.15*	0.09
aux_tense-Past	0.22*	0.15*	0.13
verb_tense-Pres	-0.25*	-0,18*	-0.06
vb_head_sent	0.12	0.004	0.26*
dep_det:poss	0.15*	0.06	0.24*
subord_dist	0.08	-0.01	0.23*
verb_form_Inf	-0.02	-0.1	0.24*
dep_advmod	-0.25*	-0.08	-0.15*
obj_pre	-0.23*	-0.003	-0.15*
princ_dist	-0.08	0.01	-0.24*

4. Conclusions

This paper has introduced a novel resource for studying and computationally modeling text coherence in the Italian language, focusing on human perception. The investigation into genre and text perturbations revealed a significant interplay between the two dimensions. Interestingly, text passages from Wikipedia were rated on average as more coherent than those extracted from TEDx talks even when presented in a perturbed form. Furthermore, a deeper analysis of the perturbations revealed distinct effects on coherence perception. Modifications that disrupted coherence by altering the sentence order or introducing intruder sentences had varying impacts. Notably, coherence judgments also varied for original texts, and the syntactic structure and complexity-related features emerged as influential factors in human assessment.

In the future we would like to gain deeper insights into the underlying factors that influence coherence perception by also incorporating a diverse range of text genres and perturbations. This deeper understanding of coherence will have significant implications for the development of more sophisticated language understanding and generation systems.

Acknowledgments

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] M. Halliday, R. Hasan, Cohesion in English, Longman Group Ltd., 1976.

[3] T. A. Van Dijk, Text and Context: Exploration in the Semantics and Pragmatics of Discourse, Longman, London, 1977.

[4] A. Lai, J. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 214–223. URL: https://aclanthology.org/W18-5023. doi:10.18653/v1/W18-5023.

[5] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 4328–4339.

[6] B. Elvevåg, P. W. Foltz, D. R. Weinberger, T. E. Goldberg, Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia, Schizophrenia research 93 (2007) 304–316.

[7] D. Iter, J. Yoon, D. Jurafsky, Automatic detection of incoherent speech for diagnosing schizophrenia, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 136–146.

[8] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, CoRR abs/2006.14799 (2020). URL: https://arxiv.org/abs/2006.14799. arXiv:2006.14799.

[9] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, J. Qi, Evaluating document coherence modeling, Transactions of the Association for Computational Linguistics 9 (2021) 621–640. URL: https://aclanthology.org/2021.tacl-1.38. doi:10.1162/tacl_a_00388.

[10] M. Chen, Z. Chu, K. Gimpel, Evaluation benchmarks and learning criteria for discourse-aware sentence representations, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 649–662. URL: https://aclanthology.org/D19-1060. doi:10.18653/v1/D19-1060.

[11] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, in: Proceedings of the First Workshop on Computational Approaches to Discourse, Association for Computational Linguistics, Online, 2020, pp. 102–112. URL: https://aclanthology.org/2020.codi-1.11. doi:10.18653/v1/2020.codi-1.11.

[12] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (1995) 203–225. URL: https://aclanthology.org/J95-2003.

[13] W. C. Mann, S. A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text 8 (1988) 243–281. URL: http://scholar.google.com/scholar.bib?q=info:BEw8CIWbucoJ:scholar.google.com/&output=citation&scisig=AAGBfm0AAAAAU3X_1Dq4ULnWfFzMeRsqGJcha1fReMSl&scisf=4&hl=en.

[14] R. Barzilay, M. Lapata, Modeling local coherence: An entity-based approach, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 141–148. URL: https://aclanthology.org/P05-1018. doi:10.3115/1219840.1219858.

[15] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 997–1006. URL: https://aclanthology.org/P11-1100.

[16] D. T. Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1320–1330. URL: https://aclanthology.org/P17-1121. doi:10.18653/v1/P17-1121.

[17] D. Brunato, F. Dell'Orletta, I. Dini, A. A. Ravelli, Coherent or not? stressing a neural language model for discourse coherence in multiple languages, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10690–10700. URL: https://aclanthology.org/2023.findings-acl.680.

[18] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computa-

tional Linguistics, Online, 2021, pp. 4164–4173. URL: https://aclanthology.org/2021.naacl-main.328. doi:10.18653/v1/2021.naacl-main.328.

[19] G. Caliendo, The popularisation of science in web-based genres, The language of popularisation: Theoretical and descriptive models 3 (2012) 101–132.

[20] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: https://aclanthology.org/2020.lrec-1.883.

[21] A. Cerulli, D. Brunato, F. Dell'Orletta, Quale testo è scritto meglio? a study on italian native speakers' perception of writing quality?, in: Proceedings of 8th Italian Conference on Computational Linguistics (CLiC-it), 26-28 January, 2022, Milan, Italy., CEUR-WS 3033, Milan, 2022, p. 9.

A. Annotation instructions

This is the questionnaire instructions provided to human raters:

"Ciao! In questo sondaggio ti chiediamo di leggere dei testi e di valutarne il livello di coerenza, assegnando un punteggio che va da 1 (per nulla coerenti) a 5 (del tutto coerenti).

Innanzitutto, ti diamo una breve definizione di coerenza: in ambito linguistico, questa parola si usa per indicare una caratteristica che riguarda l'organizzazione del significato di un testo.

Un testo è considerato coerente se le singole unità di cui si compone (tipicamente le frasi) sono connesse tra loro in modo da formare un'unità più ampia che il lettore/ascoltatore considera globalmente appropriata, sia dal punto di vista dell'ordine logico-temporale sia rispetto al contenuto principale del discorso. Tuttavia, questa valutazione è molto personale: la coerenza, infatti, dipende sia da fattori legati alla struttura linguistica e al contenuto del testo, sia da fattori soggettivi, come la familiarità del lettore/ascoltatore verso l'argomento, la sua padronanza linguistica, il grado di interesse ecc. Proprio per questo ti chiediamo di valutare ciascun testo con la maggior naturalezza possibile, dal momento che non c'è una risposta giusta o sbagliata: quello che ci interessa è proprio la tua percezione personale! In generale, per orientarti nel giudizio, puoi pensare che un testo molto coerente dovrebbe risultarti facile da comprendere, ben strutturato e non dovresti avvertire discontinuità sul piano logico e del contenuto nel passaggio tra una frase e l'altra. Ad esempio, il testo che segue dovrebbe ottenere un punteggio di 4 o 5:

E quindi che si fa? E quindi mi danno in mano un depliant dell'Università e dicono: "Bene ragazzo. Scegli una facoltà a numero aperto e, nel momento in cui qualcuno ad Economia molla, puoi subentrare te." "Benissimo!" dico. Apro il depliant dell'Università, salto a piè pari Ingegneria per l'eccessiva presenza di Chimica e tra Fisica, Filosofia, Lettere, Matematica e Informatica inizio a decidere che cosa fare.

Al contrario, un testo poco coerente dovrebbe risultarti più difficile da capire, poco coeso e discontinuo sul piano logico e strutturale. Ad esempio, il testo che segue dovrebbe ottenere un punteggio di 1 o 2:

Stiamo parlando degli anni Trenta. Le aziende, le persone che puntano al futuro, le protagoniste di questa trasformazione, non sono assolutamente associate a queste parole, semmai a: tecnologia; precisione; elettronica; digitale; meccanica; futuro. Tutte parole, queste, associate invece al termine "meccatronica". Solo il cinque per cento dei funghi che potenzialmente esistono sono stati descritti, quindi c'è veramente un mondo da scoprire sotto i nostri piedi.

Inoltre, per la tua valutazione, tieni presente che tutti i testi che leggerai non sono completi. Si tratta infatti di paragrafi di poche righe, estratti da sezioni diverse (es. introduzione, corpo, conclusione) di documenti più lunghi, che provengono da varie fonti (es. testi di Wikipedia, dialoghi trascritti). Infine, ti ricordiamo che il sondaggio è indirizzato alle persone di madrelingua italiana e la sua compilazione richiederà all'incirca 20-25 minuti.

Grazie in anticipo per la partecipazione!"

For the sake of completeness, we also report an English translation of the same guidelines:

"Hello! In this survey, we ask you to read texts and evaluate their level of coherence by assigning a score ranging from 1 (not at all coherent) to 5 (completely coherent).

Firstly, we provide a brief definition of coherence: in a linguistic context, this word is used to indicate a characteristic related to the organization of the meaning within a text.

A text is considered coherent if its individual units (typically sentences) are connected in a way that forms a broader unit that the reader/listener perceives as globally appropriate, both in terms of logical-temporal order and the main content of the discourse. However, this evaluation is highly subjective: coherence depends on factors related to the linguistic structure and content of the text, as well as subjective factors such as the reader/listener's familiarity with the topic, linguistic proficiency, level of interest, etc. That's why we ask you to assess each text as naturally as possible, as there is no right or wrong answer: what we are interested in is your personal perception! In general, to guide your judgment, you can consider that a highly coherent text should be easy to understand, well-structured, and you should not perceive any discontinuity in logical and content transitions between sentences. For example, the following text should receive a score of 4 or 5:

E quindi che si fa? E quindi mi danno in mano un depliant dell'Università e dicono: "Bene ragazzo. Scegli una facoltà a numero aperto e, nel momento in cui qualcuno ad Economia molla, puoi subentrare te." "Benissimo!" dico. Apro il depliant dell'Università, salto a piè pari Ingegneria per l'eccessiva presenza di Chimica e tra Fisica, Filosofia, Lettere, Matematica e Informatica inizio a decidere che cosa fare.

On the contrary, a text with low coherence should be more difficult for you to understand, poorly connected, and discontinuous on a logical and structural level. For example, the following text should receive a score of 1 or 2:

Stiamo parlando degli anni Trenta. Le aziende, le persone *che puntano al futuro, le protagoniste di questa trasformazione, non sono assolutamente associate a queste parole, semmai a: tecnologia; precisione; elettronica; digitale; meccanica; futuro. Tutte parole, queste, associate invece al termine "meccatronica". Solo il cinque per cento dei funghi che potenzialmente esistono sono stati descritti, quindi c'è veramente un mondo da scoprire sotto i nostri piedi.*

Furthermore, for your evaluation, please keep in mind that all the texts you will read are not complete. They are short paragraphs extracted from different sections (e.g., introduction, body, conclusion) of longer documents, coming from various sources (e.g., Wikipedia texts, transcribed dialogues). Finally, we remind you that the survey is aimed at Italian native speakers, and it should take approximately 20-25 minutes to complete.

Thank you in advance for your participation!"