# IUSSNets at DisCoTeX: A fine-tuned approach to coherence

Emma Zanoli[1], Matilde Barbini[1] and Cristiano Chesi[1]

[1]University School for Advanced Studies IUSS Pavia - NeTS Lab

**Abstract**

We present our submission to the DisCoTex shared task of the EVALITA 2023 evaluation campaign, which focuses on modeling discourse coherence for Italian texts. We highlight the importance of coherence modeling in natural language processing tasks and briefly discuss related work, including earlier linguistic theories and recent neural models. To tackle the task, we leverage pre-trained Transformer models and fine-tune them on the provided datasets. Our approach incorporates monolingual models due to limited computing resources, but shows potential for multilingual and multitask learning. Our systems ranks second overall, showing that Transformer models can be fruitfully leveraged for coherence assessment, but more work is needed to fully exploit their capabilities. The coherence assessment literature focuses primarily on English; this shared task and our work contribute to broadening the scope of current research.

**Keywords**

coherence, Transformers, NLP, computational linguistics

## 1. Introduction

Written texts are often a sequence of semantically coherent segments, designed to create a smooth transition between various subtopics [1]. Modeling coherence can be done by building text analysis models that can distinguish a coherent text from incoherent ones, or that can output a coherence score [2]. It has been a key problem in discourse analysis, with applications in many downstream NLP tasks (e.g. text generation, summarization, machine translation, dialogue generation, etc.).

Coherence modeling is at the heart of the DisCoTex shared task [3] of the EVALITA 2023 evaluation campaign [4]. This report relates the motivation and implementation of the IUSSnets team's submission.

## 2. Related work

Early computational models for text coherence assessment were mainly based on one of two linguistic theories: a) centering theory [5] and b) rhetorical structure theory [6]. In line with the first, [7] and [8] use the distribution of entity transitions over sentences to predict text coherence. In line with the second, [9] and [10] produce discourse relations over sentences with a discourse parser, showing that the relations are indicative of text coherence.

More recently, neural models have gained prominence in the task of coherence assessment. Popular examples are [11], [12], [13], and the recent state-of-the-art [14].

Our implementation choices are informed by [15], who are among the first to use Transformer models for coherence assessment.

It is interesting to note that the literature on coherence finds significant overlap with the literature on readability. The two are often likened and used as general measures of textual quality [9]. Sometimes, coherence is used as an additional feature in readability assessment [12].

By and large, the literature on automatic assessment of discourse coherence focuses on the English language. One notable exception is [16] for Danish.

## 3. Task

DisCoTEX is the first shared task focused on modelling discourse coherence for Italian real-word texts. The organizers proposed two sub-tasks:

- **Sub-task 1 - Last sentence classification**: a binary classification task. Given a short paragraph (the prompt), and an individual sentence (the target), the goal is to classify whether the target follows or not, i.e. whether joining it to the prompt gives out a coherent or incoherent text.
- **Sub-task 2 - Human score prediction**: a regression task. The goal is to predict the average coherence score assigned by human raters to short paragraphs. Judgments are expressed on a 5-point Likert Scale.

## 4. DisCoTex Data

The dataset for the DisCoTex task contains texts extracted from two sources: the Italian Wikipedia and the section of Italian speech transcripts included in the Multilingual TEDx corpus (mTEDx). For both subtasks, coherence is

*Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*

emma.zanoli@iusspavia.it (E. Zanoli);
matilde.barbini@iusspavia.it (M. Barbini);
cristiano.chesi@iusspavia.it (C. Chesi)



CEUR Workshop Proceedings (CEUR-WS.org)

analyzed within text passages of four consecutive sentences. For task 1, these were split into 8000 prompt-target pairs for each domain: the prompt is always made of the first three consecutive sentence, whereas the target can either be the actual last sentence of the passage (for the positive class) or a different one (for the negative class). This dataset is automatically generated. For task 2 there were 1064 text passages, equally balanced across the two original source datasets, of which 50% were left unaltered and 50% were artificially modified to undermine coherence. This dataset was not automatically generated: each passage was annotated by at least 10 human evaluators who were native speakers of Italian.

# 5. Description of the system

## 5.1. General intuition

For this challenge we leveraged pre-trained Transformer models and fine-tuned them on the provided data.

Transformer models [17] have been applied with tremendous success to the field of NLP. They have been shown to capture semantic relationships to a reasonable extent. As reported in Section 2, they have already successfully been applied to the task of discourse coherence modeling.

Since the DisCoTex task is tailored specifically to the Italian language, we decided to leverage monolingual Transformer models that had been pre-trained exclusively on Italian data. Given that coherence assessment datasets are available for English, we initially intended to experiment with multi-lingual transfer learning, using multilingual pre-trained Transformer models and fine-tuning them simultaneously on English and Italian data. Unfortunately, our limited computing resources did not allow us to get this far within the time frame of the shared task. Preliminary results indicate that this would have been a promising approach.

## 5.2. Pre-trained models

We experimented with 4 monolingual pre-trained models, freely available on the HuggingFace hub [18] at the time of writing:

- `bert-ita`[1]: an Italian version of BERT [19];
- `electra-ita`[2]: an Italian version of ELECTRA [19];
- `umberto`[3]: an Italian version of RoBERTa [20];
- `bertino`[4]: an Italian version of DistilBERT [21].

[1] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[2] https://huggingface.co/dbmdz/electra-base-italian-xxl-cased-discriminator
[3] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1
[4] https://huggingface.co/indigo-ai/BERTino

In the following we provide an overview of the main intuition for each model.

**BERT** by Google [22] introduced "masked language modeling" (MLM): some of the input tokens were masked, and the pre-training objective was to predict the original vocabulary id of the masked word based only on its context. MLM enabled the representation to fuse the left and the right context, leading to a bidirectional Transformer. In addition to MLM, they also used a "next sentence prediction" task that jointly pre-trained text-pair representations. After pre-training, BERT could be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, without substantial task-specific architecture modifications.

**DistilBERT** by HuggingFace [23] leveraged knowledge distillation during the pre-training phase, thus reducing the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. To leverage the inductive biases learned by larger models during pre-training, they introduced a triple loss combining language modeling, distillation and cosine-distance losses.

**RoBERTa** by Facebook AI [24] applied various pre-training enhancements to the original BERT model: longer training on longer sequences, bigger batches over more data, no next sentence prediction objective, and dynamically changing the masking pattern applied to the training data. These modifications advanced the state of the art on different downstream tasks.

**ELECTRA** by Stanford and Google [25] introduced a new pre-training task called "replaced token detection": instead of masking the input, they corrupted it by replacing some tokens with plausible alternatives sampled from a small generator network. Then, instead of predicting the original identities of the corrupted tokens, they trained a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. The model showed competitive performance compared to other models, while requiring fewer resources for training.

As previously stated, we experimented with monolingual Italian versions of these models, i.e. models that were trained using the same approaches as the ones described above, but solely on Italian data. These models were used to encode the input and return a vector representation from the last layer output (i.e. the `[CLS]` token, which was taken to signify a vector representation of the sentence).

## 5.3. Fine-tuning

The pre-trained models were fine-tuned on the available data for 10 epochs, using the following hyper-parameters: 0.1 dropout rate, 0.01 weight decay, 1e-6 learning rate, a batch size of 1 and no gradient clipping. We used the cur-

| Model | Training data | Language | Dataset size |
|---|---|---|---|
| `bert-ita` | Wikipedia, OPUS [27], OSCAR [28] | Italian | 81 GB |
| `electra-ita` | Wikipedia, OPUS [27], OSCAR [28] | Italian | 81 GB |
| `umberto` | OSCAR [28] - deduplicated | Italian | 70 GB |
| `bertino` | PAISÀ [29], ItWaC [30] | Italian | 12 GB |

**Table 1**
The pre-trained models we use and the datasets they were trained on.

| Team | Run ID | Ted | | | | | | | Wiki | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | | 1 | | | Accuracy | 0 | | | 1 | | | Accuracy |
| | | P | R | F1 | P | R | F1 | | P | R | F1 | P | R | F1 | |
| **IUSSnets** | run1 | 0.71 | 0.70 | 0.70 | 0.70 | 0.71 | 0.71 | **0.70** | — | — | — | — | — | — | — |
| | run2 | — | — | — | — | — | — | — | 0.75 | 0.71 | 0.73 | 0.72 | 0.76 | 0.74 | **0.74** |
| | run3 | 0.50 | 0.28 | 0.36 | 0.50 | 0.71 | 0.59 | 0.50 | — | — | — | — | — | — | — |
| **baseline** | Hamming | 0.51 | 0.43 | 0.47 | 0.51 | 0.59 | 0.54 | 0.51 | 0.54 | 0.50 | 0.52 | 0.54 | 0.58 | 0.56 | 0.54 |

**Table 2**
Full official results, sub-task 1. Using `bert-ita` for all 3 runs. run1: trained on the ted dataset; run2: trained on the wiki dataset; run3: trained on both datasets.

rently available PyTorch implementation of the Adam optimizer [26], `torch.optim.Adam`. During fine-tuning, the embedding layers of the pre-trained models were frozen.

### 5.4. Data

During fine-tuning, we only relied on the provided datasets. However, we used Transformer models which had been pre-trained on a variety of data sources (see Table 1).

For sub-task 2 we attempted some data augmentation techniques. Since we had a dataset were each sentence had a mean score based on at least 10 judgments, we leveraged the standard deviation to generate a distribution of 10 scores that would have the provided mean and standard deviation. We thus ended up with 10 scores for each sentence, instead of an average score. However, upon training our models on this augmented dataset, we did not notice any significant improvements and, because this approach was more resource-intensive, we eventually dropped it.

Please note that we only made use of 80% of the provided datasets during fine-tuning; the remaining 20% was used as a validation split (more details below).

## 6. Results

For the purposes of the official rankings, our results are: **0.72** on sub-task 1, **0.63** on sub-task 2.

For sub-task 1, the organizers considered the accuracy of the best run and computed the mean between the best results on the two datasets (Ted and Wiki). For sub-task 2, they first computed both Pearson and Spearman correlations, then they applied the harmonic mean between the two measures. Participants were allowed to submit up to 3 runs per sub-task; the full official results of our submission can be seen in Table 2 and Table 3.

For several months during training and experimentation, we were not made privy to the exact way the performance of our models would be calculated. The task instructions specified that task 1 would be evaluated according to accuracy and a second metric (which was never disclosed), whereas task 2 would be evaluated with a metric based on a standard correlation coefficient ("Pearson and/or Spearman" - it ended up being a harmonic mean of the two). During our experimentation, we decided to evaluate on accuracy for task 1, and on Spearman correlation for task 2. The two sections below report the respective results.

### 6.1. Sub-task 1 - evaluation results

In the absence of a test or validation set, we sampled 20% of the original training sets for preliminary evaluation. This resulted in 1600 randomly sampled data points for each dataset. On these sub-sets, we calculated the binary accuracy as implemented in the `torchmetrics` Python library[5]. We report results in Table 4.

### 6.2. Sub-task 2 - evaluation results

In the absence of a test or validation set, we sampled 20% of the original training set for preliminary evaluation. This resulted in 172 randomly sampled data points. On this sub-set, we computed the Spearman correlation coefficient as implemented in the `scipy` Python library[6]. We report results in Table 5.

[5] https://torchmetrics.readthedocs.io/en/stable/classification/accuracy.html#binaryaccuracy
[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

| Team | Run ID | Pearson Corr. | Spearman Corr. | Harm. Mean |
|---|---|---|---|---|
| **IUSSnets** | run1 | 0.50 | 0.48 | 0.49 |
| | run2 | 0.64 | 0.60 | 0.62 |
| | run3 | 0.65 | 0.62 | **0.63** |
| **baseline** | Jaccard | 0.10 | 0.13 | 0.11 |

**Table 3**
Full official results, sub-task 2. run1: `bertino`; run2: `bert-ita`; run3: `electra-ita`.

| Model | Dataset | Accuracy |
|---|---|---|
| `bert-ita` | wiki | **0.749** |
| `electra-ita` | wiki | 0.716 |
| `umberto` | wiki | 0.595 |
| `bertino` | wiki | 0.637 |
| `bert-ita` | all | **0.723** |
| `electra-ita` | all | 0.583 |
| `bert-ita` | ted | **0.704** |
| `electra-ita` | ted | 0.617 |

**Table 4**
Evaluation results, sub-task 1.

| Model | Spearman Corr. |
|---|---|
| `bert-ita` | 0.574 |
| `electra-ita` | **0.637** |
| `umberto` | 0.464 |
| `bertino` | 0.562 |

**Table 5**
Evaluation results, sub-task 2.

# 7. Discussion

The DisCoTex shared task provided us with an excellent opportunity to reflect on the notion of disourse coherence and on the ways it may be assessed, whether automatically or not.

As a preamble, let us note that datasets for coherence assessment that are automatically created by shuffling existing texts have been criticized, among others, by [31] and [32], and the models trained on them have been shown to perform weakly on downstream tasks [2]. Nonetheless, such datasets have remained common benchmarks.

Discourse coherence is a complicated concept that is related to almost every aspect of discourse communication. In the linguistics literature, there is no all-embracing rule governing coherence analysis: different scholars have presented their insight into different aspects of discourse coherence [33]. When we read a text or listen to speech, we are inclined to infuse it with coherence by making our own inferences based on our understanding and perception. Coherence is therefore achieved not by using superficial markers such as linguistic or grammatical devices, but through psychological, cognitive, or pragmatic means. The comprehension of discourse and an appreciation for its coherence are driven by active inference, background knowledge, and a degree of imagination.

It comes as no surprise, then, that the many facets of this uniquely human experience are hard to model computationally. In order to get a sense for this, we looked into the dataset collected for sub-task 2. Overall, the majority of the training dataset contained texts rated 3.0 or higher; in other words, the texts were perceived as mostly coherent. It would have been interesting to compare how the annotators rated original vs. artificially modified text passages. Although we did not have this information in the dataset, when comparing the datasets for sub-task 1 and 2, we found 19 passages in the dataset for sub-task 2 in the positive class of the ted dataset for sub-task 1: this means that these passages had not been modified from their original sources and were thus expected to be coherent. Of these 19 passages:

- none were unanimously rated as coherent, i.e. a mean score of 5 (0%);
- 4 received a mean score of 4 or above (21%);
- 10 received a rating between 3 and 4 (53%);
- 4 received a rating between 2 and 3 (21%);
- 1 even received a rating below 2 (5%).

If we were to revert these scores back to a binary classification (with a halfway cutoff at 2.5), 5 of these passages would be considered incoherent. However, for the purposes of sub-task 1, they would have been considered coherent. This simplistic example is in no way an exhaustive exploration of the nature of the tasks or the provided datasets, but it serves the purpose of reflecting on the difficulty of modeling these phenomena from a more explicit (linguistic or cognitive) perspective.

Deep learning models generally, and Transformers specifically, have been shown to capture useful semantic information in texts. Previous work has investigated Transformers for their semantic [34] and even pragmatic [35] properties. For these reasons, we hypothesized that Transformer models would be a good fit for the task of coherence assessment. Indeed, even in our simple setup, we can see promising results. Further experimentation and greater computational power could lead to significant performance improvements. Multilingual and multi-task learning might prove particularly effective in boosting performance on Italian texts by leveraging datasets that exist for the English language or for other related tasks.

Moving forward, further exploration of linguistic theories and neural models can enhance discourse coherence assessment and facilitate more sophisticated language processing applications. Focusing on more controlled textual continuations (e.g. different logical conclusions from specific premises) would shed some light on the relevance of specific factors in coherence modeling. This would also allow us to better understand the strengths and weaknesses of a transformers-based approach.

## Acknowledgments

We are thankful to the bright research community at the NeTS lab of IUSS Pavia, who encouraged and supported these experiments. This research has been partially founded by the PON Governance 2014-2020: Next Generation UPP Project - CUP D19J22000240006.

## References

[1] D. Aumiller, S. Almasian, S. Lackner, M. Gertz, Structural text segmentation of legal documents, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2–11. URL: https://doi.org/10.1145/3462757.3466085. doi:10.1145/3462757.3466085.

[2] T. Mohiuddin, P. Jwalapuram, X. Lin, S. Joty, Rethinking coherence modeling: Synthetic vs. downstream tasks, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3528–3539. URL: https://aclanthology.org/2021.eacl-main.308. doi:10.18653/v1/2021.eacl-main.308.

[3] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, DisCoTex at EVALITA 2023: Overview of the Assessing DIScourse COherence in Italian TEXts task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] 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.

[5] 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.

[6] W. C. Mann, S. A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text-interdisciplinary Journal for the Study of Discourse 8 (1988) 243–281.

[7] R. Barzilay, M. Lapata, Modeling local coherence: An entity-based approach, Computational Linguistics 34 (2008) 1–34.

[8] C. Guinaudeau, M. Strube, Graph-based local coherence modeling, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 93–103. URL: https://aclanthology.org/P13-1010.

[9] E. Pitler, A. Nenkova, Revisiting readability: A unified framework for predicting text quality, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 186–195. URL: https://aclanthology.org/D08-1020.

[10] 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.

[11] D. Tien 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.

[12] 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4328–4339. URL: https://aclanthology.org/D18-1464. doi:10.18653/v1/D18-1464.

[13] H. C. Moon, T. Mohiuddin, S. Joty, C. Xu, A unified neural coherence model, 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. 2262–2272. URL: https://aclanthology.org/D19-1231. doi:10.18653/v1/D19-1231.

[14] M. Mesgar, L. F. R. Ribeiro, I. Gurevych, A neu-

ral graph-based local coherence model, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2316–2321. URL: https://aclanthology.org/2021.findings-emnlp.199. doi:10.18653/v1/2021.findings-emnlp.199.

[15] T. Abhishek, D. Rawat, M. Gupta, V. Varma, Transformer models for text coherence assessment, 2022. arXiv:2109.02176.

[16] L. Flansmose Mikkelsen, O. Kinch, A. Jess Pedersen, O. Lacroix, DDisCo: A discourse coherence dataset for Danish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2440–2445. URL: https://aclanthology.org/2022.lrec-1.260.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, ArXiv abs/1910.03771 (2019).

[19] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[20] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[21] M. Muffo, E. Bertino, Bertino: an italian distilbert model, https://github.com/indigo-ai/BERTino, 2020.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. arXiv:1910.01108.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[25] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[26] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

[27] J. Tiedemann, Parallel data, tools and interfaces in opus, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[28] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215. doi:10.14618/ids-pub-9021.

[29] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, The PAISÀ corpus of Italian web texts, in: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 36–43. URL: https://aclanthology.org/W14-0406. doi:10.3115/v1/W14-0406.

[30] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[31] P. Laban, L. Dai, L. Bandarkar, M. A. Hearst, Can transformer models measure coherence in text: Rethinking the shuffle test, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 1058–1064. URL: https://aclanthology.org/2021.acl-short.134. doi:10.18653/v1/2021.acl-short.134.

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

[33] Y. Wang, M. Guo, A short analysis of discourse coherence, Journal of Language Teaching and Research 5 (2014) 460.

[34] E. Reif, A. Yuan, M. Wattenberg, F. B. Viegas, A. Coenen, A. Pearce, B. Kim, Visualizing and measuring the geometry of bert, Advances in Neural Information Processing Systems 32 (2019).

[35] L. Pandia, Y. Cong, A. Ettinger, Pragmatic competence of pre-trained language models through the lens of discourse connectives, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 367–379. URL: https://aclanthology.org/2021.conll-1.29. doi:10.18653/v1/2021.conll-1.29.