Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2440–2445
Marseille, 20-25 June 2022

DDisCo: A Discourse Coherence Dataset for Danish

Linea Flansmose[1]*, Oliver Kinch[2], Anders Jess Pedersen[2], Ophélie Lacroix[3]*
[1]Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus
[2]Alexandra Institute, Rued Langgaards Vej 7, 2300 Copenhagen
[3]Wunderman Thompson MAP, Glentevej 61, 2400 Copenhagen
linea.flansmose@gmail.com
{oliver.kinch, anders.j.pedersen}@alexandra.dk
ophelie.lacroix@wundermanthompson.com

Abstract

To date, there has been no resource for studying discourse coherence on real-world Danish texts. Discourse coherence has mostly been approached with the assumption that incoherent texts can be represented by coherent texts in which sentences have been shuffled. However, incoherent real-world texts rarely resemble that. We thus present DDisCo, a dataset including text from the Danish Wikipedia and Reddit annotated for *discourse coherence*. We choose to annotate real-world texts instead of relying on artificially incoherent text for training and testing models. Then, we evaluate the performance of several methods, including neural networks, on the dataset.

Keywords: discourse coherence, dataset, low-resource

1. Introduction

Coherence in text refers to the internal structure of a text as well as the overall meaning and purpose of the text in its context. This means that there are many factors to look at when deciding the overall coherence of a given text, some of which are beyond the text itself. Coherence is often described as what makes the difference between a unified text and a random selection of sentences or clauses (Halliday and Hasan, 2014). However, when people try and fail to speak or write coherently it rarely resembles a completely random selection of sentences. According to Halliday and Hasan (2014) it is the very definition of a text that it is coherent. That is, all texts written by a human should have at least some degree of coherence — but some texts can be more coherent than others.

In this study, we address two main issues in the field of discourse coherence:

1) Previous studies have mainly focused on English text and, 2) developed models using randomly shuffled texts instead of real-world annotated data.

In general, most current studies of discourse coherence modeling use highly edited and coherent data and then use a sentence ordering task to test the models. In the sentence ordering task, the model must decide whether a given text is a coherent text or an originally coherent text where the sentences have been shuffled randomly. At this point many models perform very well on this task, however, this test does not accurately reflect how humans actually write. No-one who tries but fails to write a coherent text will put sentences in a random order. Thus, these models don't measure the quality of coherence in a text but rather to what degree it resembles a human-written text or a randomly organized text. For this reason there is a need to investigate how a model would perform on a dataset of real-world texts of different qualities of coherence. This is what Lai and Tetreault (2018) set out to do and they collected and annotated a large dataset consisting of natural human-written texts with both high and low coherence. This study has set out to do the same but with Danish texts. At this stage there have been no attempt at a discourse coherence model for Danish texts and thus this paper presents an important first step for a discourse coherence model for Danish texts.

Our contribution consists in :

1. the development of the first dataset for discourse coherence in the Danish language
2. the evaluation of several methods on the dataset.

2. Previous work

At present day, there has not been many attempts at computing coherence in texts and most current studies focus on English texts.

Barzilay and Lapata (2008) made one of the first attempts at a discourse coherence model. They introduced the *Entity Grid* method which is based on centering theory and models the nouns in a sentence on a grid. The centering theory (Grosz et al., 1995) is a cognitive theory of text coherence. Here, a distinction is made between local coherence and global coherence where local coherence is at the utterance level and global coherence refers to the discourse level. In centering theory, a sentence in an utterance has centers which are semantic units. Identical sentences in two different utterances might have entirely different centers so context is important when looking for centers. Centering theory has been the basis for some computational models of coherence where centers are often simplified to

[*] Research conducted at the Alexandra Institute.

only constitute the nouns in a sentence regardless of context.

Guinaudeau and Strube (2013) expanded on the Entity Grid model and created the *Entity Graph* method which models shared nouns across sentences on a graph and is also based on centering theory.

Inspired by the concept of cohesion, (Morris and Hirst, 1991) explored lexical chains that are related. They used a thesaurus to link words that are about the same thing.

Somasundaran et al. (2014) used lexical chaining to make a model that could measure how coherent a text is in an essay-writing context. Again, they only focused on nouns although they also include adjective-noun structures.

In addition to the more traditional methods, neural models have also been used to evaluate how coherent a text is, e.g. Li and Jurafsky (2017) and Lai and Tetreault (2018). The latter compared different methods and concluded that the neural models generally performed better.

3. Dataset

3.1. Data Collection

The data collected for this project includes: blog posts from the Reddit[1] forum and encyclopedic texts from the Danish Wikipedia.[2] This data was chosen with some ideals in mind:

- the texts should be written by a variety of people;
- the texts should not be edited by professionals;
- the texts should be of a certain length;
- the dataset should ideally show texts of low, medium and high coherence;
- the data could be made publicly available under a licence that allows commercial use.

The forum blog posts from Reddit were acquired – with permission from Reddit – using the `praw` Python package, selecting comment sections from the subreddit r/Denmark[3]. The comment sections were chosen with an ideal of having at least five comments with a word length of 100-300. The encyclopedic texts from Wikipedia were extracted from the DanFEVER dataset (Nørregaard and Derczynski, 2021). The DanFEVER dataset consists of claims, Wikipedia entries, and entries from *Den Store Danske* encyclopedia[4]. We only retained the Wikipedia texts due to their length and excluded *Den Store Danske* texts as they contain more professional editing. 502 texts were chosen at random after shuffling the entire dataset. All texts were cleaned for html-tags and newlines and all texts were anonymized for names, CPR[5], phone numbers, usernames and emails using regular expressions. Only texts that contain between 100 and 300 words were retained. This cutoff was decided with inspiration from Lai and Tetreault (2018) and with the goal of having texts that were long enough to have structural coherence but not so long that annotation would be too cumbersome.

Domain	ICC	Weighted κ
Reddit	0.77	0.62
Wikipedia	0.61	0.45

Table 1: Agreement scores between the two annotators on the test set.

3.2. Data Annotation

The texts were annotated for coherence on a 3-points Likert scale:

1. low coherence
2. medium coherence
3. high coherence

Following guidelines from (Barzilay and Lapata, 2008; Burstein et al., 2013; Lai and Tetreault, 2018), texts are considered *lowly coherent* when they are difficult to understand, unorganized, contained unnecessary details and can not be summarized briefly and easily. Contrarily, texts are considered as *highly coherent* when they are easy to understand, well organized, only contain details that support the main point and can be summarized briefly and easily. A text is considered of *medium coherence* when it is relatively easy to follow, neither well nor particularly badly organized, might contain extraneous details that don't directly support the main point and might be easy enough to summarize but leave something to be desired in the structure of the text. Grammatical and typing errors are ignored (i.e. they do not affect the coherency score) and the coherence of a text is considered within its own domain.

As recommended by Lai and Tetreault (2018), we chose to leverage annotators with a strong annotation and linguistics background for annotating the texts. A subset of the data (200 texts – corresponding to the test set, see §3.3) were annotated by two annotators in order to calculate the agreement between them. We report intraclass correlation (ICC)[6] and quadratic weighted Cohen's κ scores in Table 1. The remaining data (801 texts) were annotated by one of the annotators.[7] Note that the agreement scores for the Reddit texts are much higher than for the Wikipedia texts, due to the fact that

[1] `https://www.reddit.com/`

[2] `https://da.wikipedia.org`

[3] `https://www.reddit.com/r/Denmark/`

[4] `https://denstoredanske.lex.dk/`

[5] Danish/Greenlandic Personal Identification Number

[6] We report the average raters absolute agreement (ICC1k).

[7] A linguistics expert with experience in discourse coherence and professional annotation.

Domain	Train	Test	Total
Reddit	401	100	501
Wikipedia	400	100	500
All	801	200	1001

Table 2: Number of texts in the DDisCo dataset in regards to the two domains – Wikipedia and Reddit – and the train/test split.

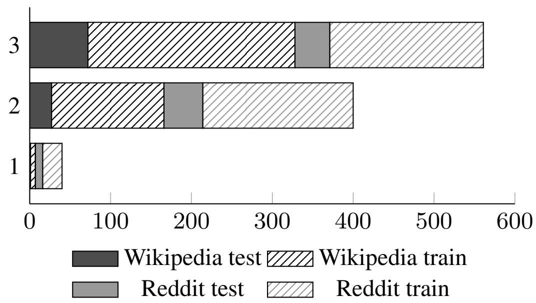

Figure 1: Distribution of coherence ratings in the dataset.

most Wikipedia texts are highly coherent and structured in a similar way which makes it difficult to differentiate from medium coherent texts.

3.3. Data Statistics

We show the amount of texts annotated for each domain and the split in test and training set in Table 2, as well as the distribution of coherence ratings in Figure 1. It is apparent that most texts across both domains received a medium or high coherence rating, and it is thus somewhat rare to find a text in the domains of encyclopedic texts and forums posts that is incoherent. There is a tendency for Wikipedia texts to have a higher coherence rating than Reddit texts.

4. Classification Methods

We compare several methods for the classification of discourse coherence rating – feature-based and text-based. As the dataset is somewhat limited in size, we perform training and evaluation on the combination of the two domains, Wikipedia and Reddit. Furthermore, according to (Lai and Tetreault, 2018), the results were better when the texts from the different domains were combined into one dataset.

4.1. Feature-based Classification

The feature-based strategy consists in pre-calculating some numerical features – relevant to discourse coherence – and using them to feed a machine learning algorithm (as single feature or in combination). We choose to compare the following four algorithms: Multinomial Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR). We describe the different features in the following paragraphs.

LIX

We choose to compute a readability mesure as a baseline feature for predicting discourse coherence ratings. The concept of coherence is often mentioned in relation to readability in texts though common readability tests do not take coherence into account.

Many factors play into readability and whether or not a text is easy or challenging to understand. In English, a very common readability test is the Flesch–Kincaid readability test (Flesch, 1979). We choose to exclude the Flesch-Kincaid mesure from our metrics, prefering one that is more adapted to Danish. In Scandinavia, a common measure of readability is *læsbarhedsindekset* (LIX) (Björnsson, 1968). However, this only takes sentence length and word length into account. It is very widely used, most likely as it is very easy to calculate and does not require much computing power. Both tests measure readability by words per sentence and syllables per word. LIX is measured by the following formula:

$$LIX = \frac{A}{B} + \frac{C * 100}{A}$$

where A is the number of words, B is the number of sentences, and C is the number of long words over 6 characters.

Entity Graph

The entity graph (Guinaudeau and Strube, 2013) is a measure of local coherence in a text. Here, coherence is measured by shared nouns between adjacent sentences. Weights are applied according to the number of shared entities as well as the syntactic role of the entity, where subjects count as 3, objects as 2 and all other nouns count as 1. The coherence score is then a measure of the average shared entities between adjacent sentences in a text.

Conjunctions

Halliday and Hasan (2014) investigated what linguistic units help make a text coherent, and called this concept cohesion. Cohesion is the study of the links between sentences or clauses that help us realize that they are related. According to Halliday and Hasan (2014), cohesion is necessary for texts: ”For a text to be coherent, it must be cohesive”. They define 5 different types of links: reference, ellipsis, substitution, lexical cohesion and conjunction.

It is difficult to measure all aspects of cohesion as it often depends on context and can be individual. Here we choose to focus on one aspect: the number of conjunctions for each text. Note that a text without conjunctions could be highly coherent, yet, we assume here that it is a predominant marker for coherence. We use part-of-speech tagging from spaCy (Honnibal and Montani, 2017) to identify both coordinating conjunctions (CCONJ) and subordinate conjunctions

(SCONJ). Both types of conjunctions were grouped as we did not expect type of conjunction to affect coherence.

4.2. Text-based Classification

In the text-based strategy, the text is directly transformed into an embedding using different pre-processing methods and then fed to a machine or deep learning algorithm for training. In all cases, neither stop words nor special characters were removed as these were expected to carry relevant information.

Machine Learning Algorithms

As previously, we choose to compare the four following algorithms for classification: Multinomial Naïve Bayes, Support Vector Machine, Random Forest and Logistic Regression. The text is pre-processed (tokenized and lemmatized) using `spaCy`[8]. We also compare two different methods for embedding the text (i.e. the entire paragraph):

- a TF-IDF vectorizer with unigrams, bigrams and trigrams;
- (Facebook) Danish word embeddings (Bojanowski et al., 2017) – learned from Wikipedia pages using FastText – averaged over all tokens in the paragraph.

Deep Learning Algorithms

We fine-tune several transformer-based pre-trained models for discourse coherence classification. Pre-processing of the text (e.g. tokenization and casing) is handled by the customed tokenizers coupled with each model, thus can differ depending on the model. We compare the following pre-trained model:

- daBERT[9] (i.e. Nordic BERT): a BERT-based (Devlin et al., 2019) model pre-trained on danish texts;
- mBERT: a multilingual BERT-based pretrained model;
- XLM-R: a multilingual XLM-Roberta-based (Conneau et al., 2020) pre-trained model.

5. Evaluation

5.1. Experiments

We report accuracy, precision, recall and weighted F_1 scores of the models on the test data.

The baseline (*Majority*) strategy represents a model that would always predict a rating of 3 (i.e. high coherence) which is the most common rating.

Each other score is an average on 5 runs. For each experiment, we split the training dataset randomly: 80% is used for training the model and 20% as development data.

Input	**Model**	**Acc.**	**Prec.**	**Rec.**	**F_1**
		Baseline			
-	Majority	0.57	0.32	0.57	0.41
		Feature-based			
LIX	RF	0.49	0.50	0.49	0.49
EGraph	RF	0.50	0.50	0.50	0.50
Conj.	RF	0.59	*0.55*	0.59	0.53
All feats	NB	*0.60*	*0.55*	*0.60*	*0.56*
		Text-based ML			
Lemmas	LR	0.58	0.33	0.58	0.42
Lemmas	SVM	0.63	0.59	0.63	*0.58*
Lemmas	NB	*0.64*	*0.61*	*0.64*	*0.58*
WV	RF	0.60	0.56	0.60	0.57
		Text-based DL (transformers)			
Text	daBERT	0.65	0.61	0.65	0.62
Text	mBERT	**0.67**	**0.64**	**0.67**	**0.63**
Text	XLM-R	0.66	0.63	0.66	**0.63**

Table 3: Discourse coherence results, i.e. accuracy (Acc.), recall (Rec.), precision (Pre.) and weighted F_1 score. Models: Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF). Inputs: Word vectors (WV). Scores in italic are the highest within the same strategy. Scores in bold are the highest globally.

For the feature-based strategy, we report only the results of the best classifier. For the text-based strategy with machine learning algorithms, we report the result of each classifier but only the one with the best text pre-processing strategy (lemmas or word embeddings).

5.2. Results

Table 3 shows the performance of the different models used for classifying discourse coherence. Globally, the deep learning models achieve the best scores.

Among the feature-based models, the conjunction feature is the most relevant for predicting discourse coherence ratings. It performs quite well even though it is very simple in nature as it simply counts the number of conjunctions in a text. This shows that conjunctions probably are an important factor in coherent texts. It could therefore be interesting to investigate the roles of other aspects of cohesion going forward. Not surprisingly, LIX is a very poor predictor of discourse coherence in Danish texts in the same way as Flesch-Kincaid is for English texts. Repetition of nouns in adjacent sentences, as seen in the Entity Graph does not appear to be much better predictors of coherent texts, which is also coherent with the study of Lai and Tetreault (2018) on English. Nevertheless, using all features combined is significantly better (+0.03 F_1) than conjunctions only.

[8] Loaded with a Danish model: `da_news_core_md`

[9] `https://github.com/botxo/nordic_bert`

For text-based machine learning algorithms, using word embeddings is not beneficial but for the Random Forest classifier. The best-performing model uses the Multinomial Naives Bayes algorithm. This model also achieves significantly better results (+0.04 acc, +0.02 F_1) than the best all-features-combined model.

As in (Lai and Tetreault, 2018) the neural models perform best.[10] The multilingual pre-trained BERT (mBERT) model achieves the best performance, though not significantly higher than the two other deep learning models.

6. Discussion

This study sets out to investigate whether a computational model could measure how coherent a natural human-written text is in the Danish Language. Compared to the current norm, we didn't use synthetic data to train models. This paper argues that the use of such a model is limited as human-written texts with low coherence do not resemble texts with random sentence order.

There are two main consequences of limiting discourse coherence studies to the evaluation of the sentence ordering task. First, the accuracy of a model trained and evaluated on synthetic data will be higher (than a model trained and evaluated on natural data) as the task is simpler. The difference between a highly coherent text and random sentence order is much bigger than that between a highly coherent text and a text with low or medium coherence. Second, some of the main potential uses often mentioned are natural language generation, document summarization, automatic essay scoring and guidance in high quality writing. Especially with the latter two, a model only tested on the sentence ordering task will likely not perform as well.

As is apparent in Figure 1, most texts were annotated with a coherence rating of 2 or 3. Only very few texts were rated as low coherence. While it is not very surprising to find a low percentage of low coherence text in general – as “all texts written by humans are coherent to some extent” (Halliday and Hasan, 2014) – there appears to be a more even distribution across the three ratings in Lai and Tetreault (2018). This bias might be mainly due to the chosen domains, which include more non-edited texts. Even though there is no official and professional editing of Wikipedia entries, it is still often written by multiple authors and an incoherent text is likely to be edited by the next author. It is thus not very surprising that Wikipedia has more texts with high coherence and almost no texts with low coherence. On another hand, the choice of the annotators also has an influence on the ratings. Although (Lai and Tetreault, 2018) recommend employing annotators “with a strong annotation background”, this idea might be questioned as they “do not reflect the demographics of end users” (Jørgensen and Søgaard, 2021) which can lead to the development of biased models and systems.

7. Conclusion

The DDisCo dataset is publicly available and can be downloaded from `https://github.com/alexandrainst/danlp`. We showed that it is possible to learn discourse coherence models on human-written Danish texts which achieve decent results – on par with recent experiments on English (Lai and Tetreault, 2018) – instead of learning from text with randomly shuffled sentences.

While this new dataset pull the Danish language away from the set of low-resource languages for discourse coherence, the size of the dataset and the domain of application are still limited. Future work include: the extension of the dataset to new domains, in particular, with a higher percentage of incoherent texts; the study of new methods for learning better models, for example in a multi-lingual setting where English data are used in combination with the Danish data; and the use of explainability methods in order to study new markers of discourse coherence.

8. Bibliographical References

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

Björnsson, C. (1968). Læsbarhed.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Burstein, J., Tetreault, J., and Chodorow, M. (2013). Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Flesch, R. (1979). How to write plain english. *University of Canterbury. Available at .[Retrieved 5 February 2016].*

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local co-

[10] In their study, they did not use text classifiers with statistical learning and thus only compared neural models to entity-based models.

herence of discourse. *Computational Linguistics*, 21(2):203–225.

Guinaudeau, C. and Strube, M. (2013). Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria, August. Association for Computational Linguistics.

Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in english*. Number 9. Routledge.

Honnibal, M. and Montani, I. (2017). SpaCy 2: Natural language understanding with bloom embeddings. *convolutional neural networks and incremental parsing*, 7(1).

Jørgensen, A. and Søgaard, A. (2021). Evaluation of summarization systems across gender, age, and race. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 51–56, Online and in Dominican Republic, November. Association for Computational Linguistics.

Lai, A. and Tetreault, J. (2018). Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia, July. Association for Computational Linguistics.

Li, J. and Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark, September. Association for Computational Linguistics.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Nørregaard, J. and Derczynski, L. (2021). DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Somasundaran, S., Burstein, J., and Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.