Topic Segmentation in Spoken Dialogue through Convergence in Utterance Complexity

Ricardo Fabián Guevara 11390786

Kwesi Menyah 11420308

Janosch Haber 10400192

Abstract

We present a simple incremental model for stable topic segmentation in spoken dialogue. The model is based on the principle of Uniform Information Density and only utilises differences in syntactic complexity of the interlocutors to predict topic shifts in upcoming turns. While this simple set of features is not sufficient to reproduce segmentations of more involved methods, we show that it produces coherent and intuitively sound topic segments even for noisy dialogue transcripts.

1 Introduction

Xu and Reitter (2018) introduced a novel, information-theoretic view of dialogue, in which they proposed modelling a conversation between two interlocutors as a two-way communication system. In such a system the information flow follows a number of general principles. One of the principles that is assumed to hold in dialogue is the Uniform Information Density hypothesis (UID, Jaeger and Levy (2007); Jaeger (2010); Temperley and Gildea (2015)). The UID hypothesis states that a communication system as a whole has the tendency to distribute data in such a way that the density of information remains constant.

In two-party dialogue both interlocutors are equal parts of the communication system. This means that they are jointly responsible for the level of information density at every moment of the conversation. In order to ensure the validity of the UID hypothesis, the two speakers must therefore have an agreement in an implicit sense about their contribution to the conversation. Xu and Reitter propose that speakers take on certain roles during a conversation: One *leads* the conversation by steering the ongoing topic, while the other *follows* along. These roles can switch during a conversation and rather than steering turn-taking behaviour, they describe a higher-level segmentation of a conversation into *topics*.

In this paper, we investigate whether we can detect the boundaries of these conversation segments, formally referred to as *topic shifts* in Conversation Analysis Keating (2000); Ng and Bradac (1993) based on the speaker's contribution to the conversation alone. To this end we extract a number of simple syntactical features that have been shown to correlate well with the amount of information transmitted (Genzel and Charniak (2002); Jaeger and Levy (2007); Jaeger (2010)) and build a simple prediction model based on these features. While we had to conclude that this simple approach does not yield a model expressive enough to correctly predict topic shifts produced by more involved methods, we claim that it nonetheless produces coherent and intuitively sound topic segments even for noisy dialogue transcripts.

In Section 2 we present previous work on topic segment analysis. Section 3 then introduces our approach to topic shift prediction, describing the data, features and models used. Due to negative results, we propose a second approach detailed in 4 that makes further use of the collected data. Section 5 concludes this paper with a summary of our conclusions and a discussion of our findings.

2 Related Work

Conversation Analysis theories state that regular, non task-driven conversations contain several topic episodes (Keating (2000); Ng and Bradac (1993)). Within each of these episodes, one of the speakers will take on a leading role, unfolding a new topic, while others play a more passive role and follow the topic shift. Introducing a new topic thus means that the topic leader supplies a large

contribution of new information, whereas towards the end of a topic segment neither of the speakers would have any more relevant information to contribute to the topic. We can therefore assume that the information contributions from each interlocutor converge in a topic segment.

2.1 Measures of Information Content

Xu and Reitter (2018) analyse these patterns of information contribution to a conversation through an information-theoretic model, formulating sentence information H(S) as

$$H(S) = H(w_1...w_n) \quad (1)$$

$$\approx -\frac{1}{n}\sum_{w_i \in W} \log P(w_i|w_1....w_{i-1}) \quad (2)$$

$$\approx -\frac{1}{n}\sum_{w_i \in W} \log P(w_i|w_{i-2}, w_{i-1}) \quad (3)$$

using a trigram language model, where $P(w_i|w_{i-2}, w_{i-1})$ is estimated through Katz backoff Katz (1987) (see B for more details). They apply this measure to topic segments produced by TextTiling (Hearst, 1997), one of the most well-known topic segmentation algorithms for written text, since no gold-standard annotated data is available at the present time. TextTiling determines topic shifts based on lexical similarity between consecutive sequences of words. It splits the text into sequences of words of fixed length, then measures the similarity between adjacent sequences and places boundaries whenever there is a radical change in this measure. The boundaries are then moved to the closest paragraph change in the original text.

2.2 Assignment of Speaker Roles

Having segmented all conversations into topic episodes with TextTiling, the next step in calculating the development of sentence information in the segments is to determine which speaker takes on the role of topic leader. To do so, Xu and Reitter propose two rules for speaker role differentiation:

- **Rule 1:** If the topic episode starts in the middle of the speaking turn of a given speaker, let that speaker be the leader of this topic segment.
- **Rule 2:** If the topic episode starts with a new speaking turn, let the first speaker who contributes a sentence longer than five words be the leader of this segment.

Dividing the speakers in this way and collecting sentence information for each role individually, Xu and Reitter observe the expected convergence of contribution within a topic segment.

3 Topic Shift Prediction

Assuming that convergence within topic segments exhibits stable behaviour, we propose to reverse the approach of Xu and Reitter and use the convergence of contribution as a feature to predict topic segments. More precisely, we propose an incremental model that, given appropriate features from an ongoing topic segment, predicts after each utterance whether a new topic will be introduced in the next utterance. This way the model can readily be used in real-time spoken dialogue segmentation and can be introduced as a module in artificial dialogue agents to determine when to introduce new information into a conversation - and when to follow the other speaker in an ongoing topic. To do so, the proposed model must learn the correct weights of a set of sentence features so that the combination of activations induced by a given sentence can be used to determine whether the next utterance is likely to belong to a new topic or will be a continuation of the current one.

3.1 Data

For all experiments in this paper we use the Switchboard SwDA corpus (Godfrey et al., 1992). We also did the convergence experiments with the British National Corpus (BNC, BNC Consortium (2007)) to verify if the same behaviour is observed. Albeit present, it is more noisy and so any other experiments were performed on the SwDA only. SwDA contains 641 telephone conversations of pairs of US-American English native speakers asked to discuss a given topic. From the BNC we use the BNC-DEM section (spoken, casual conversations) that contains 1352 conversations, filtered down to 1290 2-party conversations, which we pre-processed with the Stanford parser (Chen and Manning, 2014) to obtain sentence syntax data. A detailed summary of the two data sets is displayed in table 1.

3.2 Features

Due to the incremental nature of the model, we only want to use features that can be calculated locally. Sentence information however can be ex-

Dataset	Size	Filters	Syntax data
BNC-DEM	1290	2 speakers	Stanford parser
SwDA	641	-	Included

Table 1: Details of the two data sets used in this study.

pensive to calculate and requires a pre-processing of the entire corpus to obtain global frequency counts. We therefore require a simpler set of features to approximate sentence information measures in our model.

Previous studies have pointed out that sentence information is closely correlated with other syntactic complexity measures of sentence (Genzel and Charniak (2002); Jaeger and Levy (2007); Jaeger (2010)). Like Xu and Reitter (2016) we follow this approach and analyse speaker contributions using three simple, syntax-related features: Sentence length, tree depth and tree width.

1. **Sentence length (SL)**: Number of words. Longer sentences are likely to indicate more sentence information

2. **Tree depth (TD)**: Longest distance between root and leaf nodes in the parse tree of a sentence

3. **Tree width (TW)**: Average number of children of all non-leaf nodes in the parse tree of a sentence. Syntactically complex sentences are typically used to express more complex meanings and are therefore likely to contain more information

Both tree depth and tree width are strongly correlated with sentence length and should be normalised. Following Xu and Reitter, we calculate the average respective measure for sentences of the same length and use it as normalising constant. This way we obtain two additional features:

4. **Normalised tree depth (NTD)**: Tree depth normalised by average depth per sentence length

5. **Normalised tree width (NTW)**: Tree width normalised by average width per sentence length

Figure 1 shows the average development of these features within topic segments in a two-party dialogue as produced by TextTiling on the SwDA data set. Whilst we see similar convergence behaviour for the non-normalised measures as observed by Xu and Reitter, with our implementation we also obtain fairly stable behaviour for the normalised measures. Here Xu and Reitter reported a much smaller although significant convergence behaviour on Switchboard - but had interchanged contributions for the leader and follower roles in their NTW.

Having collected the five complexity measures for each utterance in the two data sets, we annotate each utterance with a binary label: 1 if the next utterance is the initial utterance of a new topic segment as predicted by TextTiling and 0 otherwise. In this way we obtain an annotated dataset that can be used to learn optimal feature weights through supervised machine learning.

3.2.1 Additional Features

Given that having only a handful of non-independent features may be a limiting factor for successful training, we propose calculating in addition a number of higher-level, relative features using the following basic metrics as a starting point:

6. **Initial/previous/current difference**: Calculates the absolute difference between the two speaker's contributions during the first/previous/current turn of the topic episode.

9. **Contribution development**: Calculates the degree of change between the initial and current difference in speaker contribution through dividing the current difference by the initial difference. This can also be done for the development from previous to current turn.

11. **Utterance counter**: Indicates the utterance's position within the current topic.

12. **Speaker role**: 1 if the speaker is the topic leader, 0 otherwise.

Each row in the resulting training dataset contains 42 features: the five basic measures for the initial, previous and current utterance (15), the difference in contribution for the basic measures at those three points in time (15), the ratio of change between initial and current and previous and current (10), the utterance counter and the binary role label.

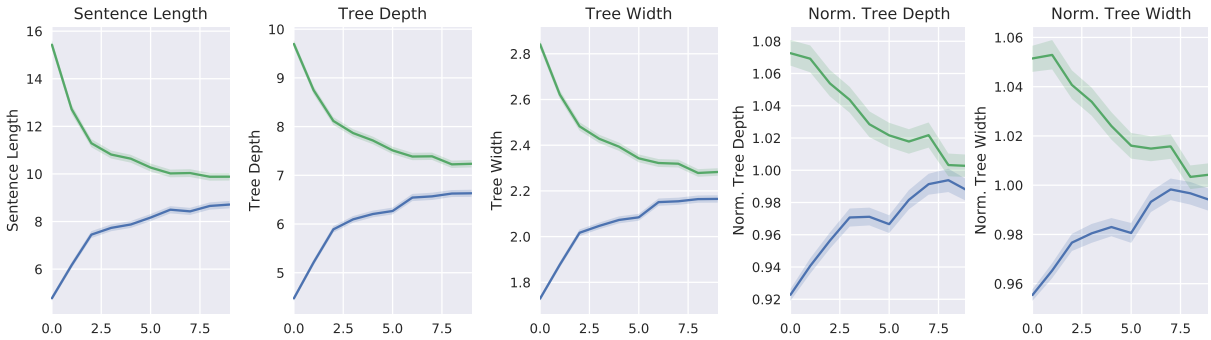

Figure 1: Development of the five basic features within topic shifts produced by TextTiling on the SwDA data set for topic leader (green) and follower (blue) roles. The x axis ranges over the position of an utterance within a topic.

3.3 Models

Now that we have a satisfactory amount of features, the next issue to address is the skewed distribution of labels: Naturally the 'shift' category will be significantly less frequent than the 'non-shift' one, resulting in a ratio of about 33:1 in SwDA. Since this unbalanced distribution of data hampers naive approaches to Machine Learning, we propose to use two different models that have shown good performance in dealing with skewed data: An auto-encoder (AE) outlier detection model and a simple multilayer perceptron (MLP) binary classifier with oversampling.

3.3.1 Auto-encoder Outlier Detection

The reason for using an auto-encoder to address this problem is that we can filter out all sentences labelled as inducing topic shifts from the training data and have the auto-encoder optimise its parameters so as to optimally reconstruct this filtered dataset. Its performance in this task is measured through a reconstruction loss: The closer the produced outputs are to the actual inputs of the model, the lower the loss. So if there are underlying characteristics describing sentences which are not followed by a topic shift in our data, the model attempts to learn those characteristics to improve its performance. Consequently, assuming that utterances which induce a topic shift have different characteristics, this will mean that their encodings will result in a higher reconstruction loss as the model cannot fit them to the learned characteristics. We can thus use a difference in reconstruction loss to classify novel sentences into two categories: Shift and non-shift utterances.

3.3.2 Multilayer Perceptron (MLP) Binary Classifier

MLPs on the other hand are one of the most widely applicable machine learning tools, optimising their parameters to reduce classification error on a binary annotated training set. This simple model however performs badly on highly skewed data: if it were to predict every sentence to be non-shift, this would still yield a near perfect accuracy of 96% - while at the same time not detecting any of the topic shifts. To counter this issue, a technique called oversampling can be used: Every training batch is enriched with an artificially high number of shift samples so that classifying them correctly becomes equally important to the model.

3.4 Experiments

For the auto-encoder, we use a simple architecture that was shown to work well for detection of credit card fraud on the Kaggle dataset, as implemented by Venelin Valkov. We train until convergence at about 50 epochs with a batch size of 32, using a 80-20 train-test split of our dataset where we remove all topic shift samples from the training set. We therefore obtain 90855 training samples and 23433 test samples. The data is normalised to zero mean.

For the MLP, we implemented two hidden layers with 20 nodes each using ReLU activations. The learning rate was set to a constant 0.001. We train until convergence at about 100 epochs with a batch size of 64, using the same dataset as before but instead of removing topic shift samples from the train set we use SMOTE oversampling (Chawla et al., 2002) to produce artificial sam-

ples close to the underrepresented shift samples. We therefore obtain 181710 training samples and 23433 test samples.

3.5 Results

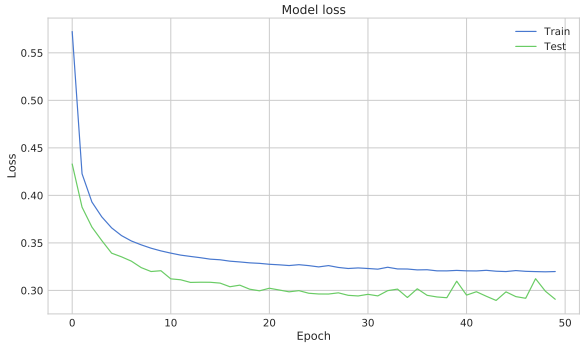

Figure 2: Train and test set error of the auto-encoder. The train set does not contain any topic shift samples.

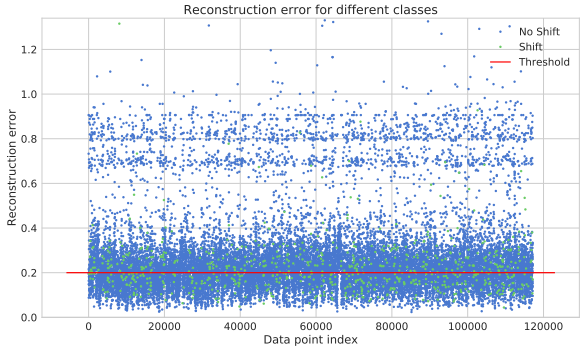

Figure 3: Reconstruction errors for all data samples. The threshold line should split topic shift samples (green) from non-shift samples (blue).

Figure 2 shows the training and test loss of the auto-encoder outlier detection model. The average loss on the test set here is even lower than the training set error, indicating that the training set contains a large number of outliers that do not actually indicate topic shifts. Since our goal is to use the reconstruction error to detect samples that indicate topic shifts, we therefore expect a high number of false positives when we try to split the data samples into shift and non-shift groups. This is validated by plotting the per-sample reconstruction error in Figure 3: The objective is to place the horizontal threshold line so that it separates shift and non-shift samples as good as possible - but it is not possible to actually perform such a split since shift and non-shift samples seem to be equally distributed, i.e. they do not produce significantly different reconstruction errors. We therefore cannot use this method to learn optimal feature weights for the topic shift prediction task.

When training the MLP using SMOTE oversampling with ratio 1:1, we obtain about 32% test set accuracy - with still only 3 correct predictions of topic shifts and more than 6000 wrongly predicted shifts.

3.6 Conclusion

We appear to be unable to predict the topic shifts produced by TextTiling with the provided syntactical complexity features alone. Since intuitively the clear convergence behaviour observed within an average topic should provide enough information to reproduce the topic shifts, we find this a surprising result.

Our main hypothesis to explain why the local syntax features could be insufficient to predict TextTiling's topics is that TextTiling itself uses more global information to segment a conversation. Topic shifts therefore do not only depend on previous utterances, but also on subsequent utterances. The observed convergence behaviour therefore might be an effect that correlates with TextTiling segmentation of topics - but might not be causally related to them. To validate this hypothesis, we collected the same metrics as before for randomly inserted topic shifts that produce about the same number of topics as TextTiling. Figure 4 shows the average behaviour of sentence length within those random segments: We can observe a significant convergence behaviour even here. Convergence of syntactic complexity alone therefore appears not to be sufficient for predicting TextTiling topics.

4 Topic Segmentation

While syntactical convergence might not be sufficient to predict other model's segmentations, it can still be used to produce topic segments itself. We therefore propose a different approach to topic segmentation with the available syntax features: Using a simple threshold model on the activation of the collected features, we produce our own segmentation of the data and assess its quality. Since here again no topic-annotated gold-standard data

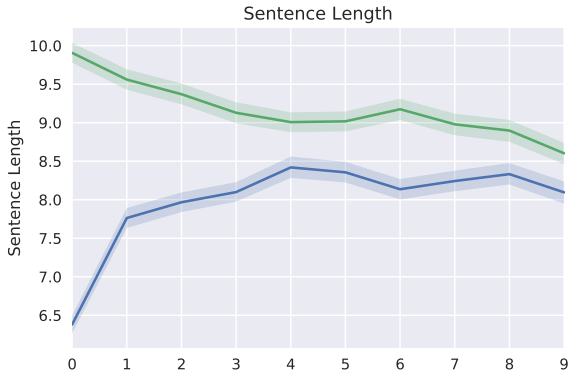

Figure 4: Development of sentence length for topic leader (green) and follower (blue) within random segments of the SwDA data set.

set is available to compare our results to, we propose to use a set of external measures to compare our model to other segmentation methods, like e.g. TextTiling, and report the relative level of performance.

4.1 Model

For our simple syntactical topic segmentation model we start with a baseline that only utilises the most basic form of local relative features. We propose this baseline model to work as follows: Determine the speaker roles at the onset of a new topic. Then iteratively parse any succeeding utterance $u_i \in T$ and record the five basic complexity measures for u_i. As soon as a minimum of two measures are obtained for each speaker, we calculate the difference in contribution in the initial utterance u_0 (feature 6) and current utterance u_i (feature 7) and determine the development in those measures between the start of the topic and the current utterance (feature 9). This returns an array of five ratios which increase when speaker contributions converge. Subsequently, we take a weighted sum of these five ratios and check if the result passes a given threshold (see Appendix C for a schematic view).
Using this model setup we obtain a total of six parameters to be optimised: The five feature weights and the threshold. Due to the threshold setup we however can no longer automatically learn the optimal parameters and need to run a grid search pipeline to determine them based on a given evaluation metric instead.

4.2 Evaluation

In order to quantitatively evaluate model performance, optimise its parameters - and compare its outputs to those of other segmentation methods - we need a set of metrics that capture the *cohesion* of the produced topic segments. Cohesion indicates how separate a predicted topic is from the next. If there's too much cohesion between contiguous topics, then it is likely that they should have been predicted as a single topic instead. In order to obtain these metrics, we use a subset of the measures provided by TAACO, the tool for the automatic analysis of text cohesion (Crossley et al., 2016):

1. **Adjacent paragraph overlap of content words (AP Content Overlap)**: This metric measures how much content is shared between contiguous paragraphs. In our usage of the tool, a paragraph equals a topic. Specifically, this metric reports the average number of words in scope that each topic have in common with the next topic. For this, the text is lemmatized (e.g. verbs are all transformed to infinitve form, nouns are all turned to singular form) before doing the calculation. The total number of overlap words for all topics, is divided by the total number of words in the whole text. Only content words are considered.

2. **Adjacent paragraph overlap of content words, paragraph normed (Paragraph-normalised APCO)**: This metric is similar to Adjacent paragraph overlap of content words, but here instead of dividing by the total number of words in the text, the denominator is now the number of topics. So while the first metric can be interpreted as a proportion of repetition amongst consecutive paragraphs, this metric reports average number of content words each pair of contiguous topics repeat.

4.3 Experiments

In order to evaluate our model performance, we compared its scores on the AP Content Overlap and Paragraph-normalised APCO metrics under different parameter combinations to those of TextTiling and the random segmentation with an average of 12 topics per conversation. Since equally increasing weights and decreasing the threshold has the same effect, we fixed the threshold to

Model	Parameters	No. Topics	AP Content Overlap	Par.-norm. APCO
TextTiling	-	7.436	0.339	11.464
Random	-	11.587	**0.201****	7.247**
Threshold = 5	[1, 0.25, 0.25, 1, 1]	12.74	0.231**	8.884**
Threshold = 5	[3, 0, 0, 0, 0]	19.916	0.212**	**5.731****

Table 2: Excerpt from the grid search for optimal parameters of the simple threshold topic segmentation model. Scores marked ** are significantly different to the TextTiling results ($p << 0.05$)

an arbitrary set value (=5) and only adjusted the weights accordingly.

4.4 Results

For an overview of some of the tested combinations see Table 2. After an initial series of supervised runs we observed that sentence length alone often already determines whether the threshold is passed or not. Only considering sentence length therefore did not change the outputs substantially, and also gave us the best results on the evaluated metrics. After some qualitative evaluation we however only accredit a very limited expressiveness to these metrics; especially since random topic segments consistently and significantly outperform TextTiling and the topic segments proposed by our model for almost all of the possible parameter combinations.

4.5 Qualitative Analysis of produced samples

The intuitively best topics are produced by our model with the threshold set to 5 and using the feature weight vector [1, 0.25, 0.25, 1, 1]. An excerpt from a sample output of this model configuration is displayed in AppendixD. The two speakers were asked to talk about what they wear at work. Here we will highlight some of the reasons why we think that the topics displayed are intuitively 'good'. Notice that we have no objective measure for this criterion.

1. The second topic covers speaker A's explanation about her style of clothing at work and ends with her asking 'How about you?', changing the focus of the conversation. In the third topic, speaker B then explains her style of work wear.

2. Topic 5 starts with an elaborated statement from speaker B who finds mini skirts quite un-professional. Toward the end, speaker A uses the 'there are mini skirts and there are mini skirts' argument to start a new topic. Topic 6 then starts with speaker A's opinion on mini skirts and starts a somewhat longer discussion.

3. In topic 7 speaker B then introduces a new argument about feeling comfortable with one's choice of clothing based on your environment. In the same manner, most of the following topics are introduced by a longer statement of one of the speakers followed by some form of discussion based on it.

4. The last two topics of the conversation then nicely cover the formal ending and greeting part

Neither the topics produced by TextTiling (in this case only two) nor the random topic segments exhibit any of the reported markers. Outputs of our model with other parameters only exhibit some of the features outlined above - but still a large part of them scores better on the cohesion metrics than the model that produced the displayed output. We therefore suggest that neither the topics produced by TextTiling nor the cohesion metrics actually capture what we would intuitively call a topic in spoken dialogue.

5 Conclusion

In this paper we show that while observing a clear convergence of speaker contribution in a dialogue topic segment as measured by a number of measures capturing syntactic complexity, we appear to be unable to predict the topic segments using these features from preceding utterances only. We can however utilise the convergence behaviour to incrementally produce new topic segments with very limited computational cost. The resulting topic segments exhibit convergence of speaker contribution like those of other models and seem to better capture what we would intuitively call a topic in a common conversation. The unavailability of annotated test data however

precludes an objective evaluation of these results and should be the focus of succeeding research.

Although we can largely reproduce the convergence behaviour as identified by Xu and Reitter (2016) in TextTiling topics, we also observe comparable convergence patterns in random text segments. We therefore question the existence of a strong relationship between what TextTiling labels as topics and the re-occuring syntactical convergence of speaker contributions in a dialogue. We rather propose that it might be a fundamental characteristic of spoken conversations that new information is added whenever both speakers can assume that previous information was taken in by both speakers and is agreed upon. The task of subsequent utterances then is to make sure that this new information is shared by the speakers, introducing less and less additional information until convergence. Defining these exchange patterns as topics, our model provides a very simple way to incrementally predict when speakers have reached agreement about the information shared and thus new information should be added. While we currently cannot yet test this against a gold standard model of topic segmentation of spoken dialogue, we strongly encourage further research into this novel approach.

References

BNC Consortium. 2007. The british national corpus, version 3 (bnc xml edition). Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16(1):321–357. http://dl.acm.org/citation.cfm?id=1622407.1622416.

Dangi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. *Proceedings of EMNLP 2014* .

Scott. A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48(4):1227–1237. https://doi.org/10.3758/s13428-015-0651-7.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 199–206. https://doi.org/10.3115/1073083.1073117.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*. IEEE Computer Society, Washington, DC, USA, ICASSP'92, pages 517–520. http://dl.acm.org/citation.cfm?id=1895550.1895693.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.* 23(1):33–64.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1):23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002.

T. Florian. Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, MIT Press, pages 849–856.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35(3):400–401. https://doi.org/10.1109/TASSP.1987.1165125.

Elizabeth Keating. 2000. Per linell, approaching dialogue: Talk, interaction and contexts in dialogical perspectives. (impact: Studies in language and society, 3.) amsterdam & philadelphia: Benjamins, 1998. pp. xvii, 330. *Language in Society* 29(4):586589.

Sik Hung Ng and James J. Bradac. 1993. *Power in language: Verbal communication and social influence.*. Sage Publications, Inc, Thousand Oaks, CA, US.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

David Temperley and Daniel Gildea. 2015. Information density and syntactic repetition. *Cognitive Science* 39(8):1802–1823. https://doi.org/10.1111/cogs.12215.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* page 443448.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition* 170(Supplement C):147 – 163. https://doi.org/10.1016/j.cognition.2017.09.018.

A Contribution

Janosch collected the features and implemented, trained and optimised the models. Fabián prepared the BNC corpus and took care of the evaluation of model outputs using TAACO. Kwesi contributed to the project and proofread the final report.

B Information Metrics

Xu and Reitter (2018) analyse these patterns of information contribution to a conversation through an information-theoretic model. Based on the formulation of entropy:

$$\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\log(\mathrm{P}(X))] \quad (4)$$

$$= \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) \quad (5)$$

$$= -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i) \quad (6)$$

by Shannon (1948), which originally is a measure of the capacity of a channel, they formulate sentence information $\mathrm{H}(S)$ as:

$$H(S) = H(w_1 ... w_n) \quad (7)$$

$$\approx -\frac{1}{n} \sum_{w_i \in W} \log P(w_i | w_1 .... w_{i-1}) \quad (8)$$

$$\approx -\frac{1}{n} \sum_{w_i \in W} \log P(w_i | w_{i-2}, w_{i-1}) \quad (9)$$

using a trigram language model, where $P(w_i|w_{i-2}, w_{i-1})$ is estimated through Katz backoff Katz (1987). Since this however strongly correlates with sentence length, following an approach by Genzel and Charniak (2002) they further propose to normalise sentence information by dividing a sentence's information content by the average information content of sentences of the same length. To do this, they calculate the average information $H(n)$ for all sentences of length n:

$$H(n) = \frac{1}{|S(n)|} \sum_{s \in L(n)} H(s) \quad (10)$$

and normalised sentence information $H'(s)$ consequently as:

$$H'(s) = \frac{H(s)}{H(n)} \quad (11)$$

With this measure, they observe that sentence information significantly increases with the global position of an utterance - which is a well-known phenomenon from the domain of written text analysis.

C Threshold Topic Segmentation Model

Data: transcript T, weights W
Result: Next topic shift index
leader measures L = [0,0,0,0,0];
follower measures F = [0,0,0,0,0];
while *T contains utterances* **do**
  U = next utterance in T ;
  update measures(U); **if** *(length(leader measures)* $\geq$ *2 & length(follower measures)* $\geq$ *2)* **then**
    initial distance I = L[0] - F[0];
    current distance C = L[-1] - F[-1];
    development D = (I - C) / I;
    topic score = weighted sum(D, W);
    if *(topic score* $>$ *threshold)* **then**
      return index;
    end
  end
end

Algorithm 1: Simple threshold topic segmentation algorithm

D Example Model Output

—— Topic 1 start at 0 ——
B: Okay, B: what do you usually wear to work? A: Well, uh, I am basically retired now. B: Uh-huh.

—— Topic 2 start at 4 ——
A: I was a member, A: I was in education and in administration, B: Uh-huh. B: Uh-huh. A: And, uh, heels, A: and I was never one, uh, because my work often took me into court, uh, never was one that got, uh, accustomed to wearing pants suits and pants to work. B: Uh-huh. A: But that was just me. A: I know many people are co, very comfortable in the classroom and what have you wearing pants. A: Uh, it, A: I guess I was just old enough not to, uh, be very comfortable in it. B: Uh-huh. A: How about you?

—— Topic 3 start at 17 ——
B: Well, I work at T I B: and they do n't really have, uh, dress code, so to speak there. B: It 's

Figure 5: Development of the five basic features within topic shifts produced by TextTiling on the BNC data set for topic leader (green) and follower (blue) roles. The x axis ranges over the position of an utterance within a topic.

pretty lax about, um, you know, B: we-, you can pretty much wear whatever you want to, B: and I wear anything from jeans, when I 'm feeling really casual to, uh, suits and dresses when I 'm meeting with a customer A: Oh. B: and so when I 'm teaching a class, obviously I wear a suit or dress, A: Uh-huh. B: It, it, uh, definitely fluctuates mainly with what I 'm going to be doing that day and kind of what my mood is B: and when it 's raining, B: I 'm more likely to wear jeans B: and, and when it 's really cold I 'm more likely to wear jeans or pants or sweaters, or that type of thing. [...]

—— Topic 7 start at 86 ——
B: Well I feel like too, on the job, when, you know, there 's men around and some of the managers are men, you just, you know, you do n't want them looking at your legs necessarily. A: That 's right. B: And, uh, to me I just would n't feel comfortable in that at work, B: but, uh, A: Well, I, I, uh, I have to, uh, agree with that, even when they was very, very popular in the early sixties, B: Uh-huh. A: uh, I, uh, uh,

—— Topic 8 start at 93 ——
A: again maybe because I was at the school, there were still many teachers who wore mini skirts, B: Uh-huh. A: uh, we had no regulation against it A: and a lot of the kids did of course, A: and it could be very embarrassing for the men teachers. B: Right. A: Because they were not that careful in how they handled themselves in those mini skirts. B: Right. [...] A: Right, A: right. B: So, that 's I think, that is good that they 're like that. B: I do know there 's a lot of companies that are very strict about what the employees wear B: and they must wear blue or gray or black and a white shirt and, A: Yeah. B: you know, no variation, B: and I do n't, I do n't quite agree with that. A: Yeah, A: well, I do n't either.

—— Topic 11 start at 174 ——
A: Fortunately I do n't have to work in those companies. B: Right. A: But, uh, I, I, uh, did have a group come over from one of the banks, over the children 's hospital where I was volunteering A: and, uh, they were doing a presentation A: and every one of the young execs coming up were dressed exactly a like, men and women. A: They all had on the gray jackets and the gray trousers or sla-, or skirts and the white blouses and the same color tie. B: Uh, right. A: And one was a skirt and one was a pant. B: Yeah. A: And, uh, and I think that 's sad because that does n't allow for any individuality. B: Uh-huh. A: That 's, uh, can stifle creativity. B: Right, B: I agree. A: So. B: Well, it,

—— Topic 12 start at 190 ——
B: I guess we 've talked probably long enough. A: I guess so, A: well it 's been nice talking with you. B: Nice talking to you too, B: I enjoyed it. A: Uh-huh,

—— Topic 13 start at 196 ——
A: bye, bye. B: Bye, bye.