A Psychology-Driven Computational Analysis of Political Interviews

Darren Cook[1], *Miri Zilka*[2], *Simon Maskell*[1], *Laurence Alison*[1]

[1]University of Liverpool, UK
[2]University of Sussex, UK

{d.cook, smaskell, alisonl}@liverpool.ac.uk, m.zilka@sussex.ac.uk

Abstract

Can an interviewer influence the cooperativeness of an interviewee? The role of an interviewer in actualising a successful interview is an active field of social psychological research. A large-scale analysis of interviews, however, typically involves time-exorbitant manual tasks and considerable human effort. Despite recent advances in computational fields, many automated methods continue to rely on manually labelled training data to establish ground-truth. This reliance obscures explainability and hinders the mobility of analysis between applications. In this work, we introduce a cross-disciplinary approach to analysing interviewer efficacy. We suggest computational success measures as a transparent, automated, and reproducible alternative for pre-labelled data. We validate these measures with a small-scale study with human-responders. To study the interviewer's influence on the interviewee we utilise features informed by social psychological theory to predict interview quality based on the interviewer's linguistic behaviour. Our psychologically informed model significantly outperforms a bag-of-words model, demonstrating the strength of a cross-disciplinary approach toward the analysis of conversational data at scale.

Index Terms: conversation analysis, automation, interviewing, communication accommodation theory

1. Introduction

Analysing interviews to deduce how an interviewer may steer an interviewee in the desired direction is not a new endeavour. Harnessing the computational power of machine learning in favour of this undertaking is a more recent development. Traditionally, analysis of conversation focused on qualitative study and required considerable manual effort, limiting the volume of analysed data. Offsetting time-exorbitant tasks to a computer presents the promise of scalability, alongside reproducibility and, debatably, greater objectivity.

Conversational outcomes have been computationally evaluated in domains such as group collaboration [1], job interviewing [2], speed-dating [3], hostage negotiation [4], and police interrogation [5]. Typically, computational models rely on manually pre-labelled data to separate the successful interactions from the non-successful ones. This reliance gives rise to potential limitations. Notably, the success measure is often not well defined, hindering the transparency of the model and as a result, its mobility between applications. In addition, domain-specific knowledge may be encoded within the labelled data, but it is not decoded into the model itself, limiting the utility of the model and any conclusion drawn with respect to the domain of origin.

To address these issues, we must develop and encourage the use of cross-disciplinary approaches. Ideally, we can capitalise on domain knowledge and expertise while exploiting the utility of computation. This paper aims to demonstrate the power of cross-disciplinary analysis by employing social psychological insight to engineer automated features and outcome measures for the benefit of interview analysis. In this paper, we demonstrate our approach on political interviews. Important in their own right, political interviews make an excellent model for conversations where the objectives of the interviewer may be in tension with the objective of the interviewee.

This paper seeks to answer two main questions: (1) can we predict, computationally, if a human will regard an interview as successful without relying on pre-labelled data? (2) Can we predict the success of an interview from the behaviour of the interviewer? The latter question originates from existing social psychological theories [6]. Specifically, communication accommodation theory [7] postulates that speakers modulate their linguistic behaviour relative to one another to meet particular objectives. Research has demonstrated that increased accommodation activity contributes to social effects such as personal attractiveness [8], and conversational fluidity [9].

Our contributions are as follows: A flexible framework for computationally analysing success in non-labeled interview transcripts, validated by a small-scale study with human participants. A set of engineered features, informed by social scientific theory, used to predict success in an interview based on the behaviour of the interviewer. We demonstrate that these domain-informed features significantly outperform a bag-of-words model. In addition, our results feed back to domain knowledge, corroborating that interviews operate in a manner consistent with less-adversarial forms of conversation.

2. Methods

2.1. Corpus pre-processing

2.1.1. Dataset

We created a corpus of N=684 political interviews from transcripts of six US cable news networks (CNN, NBC, ABC, MSNBC, CBS and Fox). We included interview segments comprising a single interviewer and a sole interviewee. We considered an interview *political* if the interviewee was a government member or had a clear affiliation with a political party. Interviews were conducted between 2013-2020 and featured 261 participants (55 interviewers and 206 interviewees). Transcripts were opportunistically-sampled from online repositories, stored as plain text files, and spot-checked to ensure accuracy. The length of interviews varied between 549 and 9102 words (M=1883.3, SD=1113.35). In total, the corpus comprised just under 1.3m words and 28,022 speech turns (M=40.97, SD=47.46).

2.1.2. Data cleaning

The python script produced to clean and standardise the transcripts is available on GitHub under an MIT licence. All non-ASCII characters and punctuation symbols were removed

Table 1: *Sample of pre-processed text. NG, LC and DR refer to n-gram, lexical category, and dependency relation models, respectively.*

Model	Example
NG	"The" (unigram), "Quick" (unigram), "Brown Fox" (bigram), "Jumps Over" (bigram), "The Lazy Dog" (trigram)
LC	"The" (article), "Quick" (relativity), "Brown" (perception), "Jumps" (motion), "Over" (power), "The" (article), "Lazy" (neg_emotion)
DR	"The" (determiner), "Quick" (adjectival modifier), "Brown" (adjectival modifier), "Fox" (nominal subject), "Jumps" (ROOT), "Over" (preposition), "The" (determiner), "Lazy" (adjectival modifier), "Dog" (preposition object)

along with timestamps, annotations, and prerecorded segments. We also expanded contracted words using the dictionary provided in [10]. Transcripts were split into a sequence of speech turns. Each speech turn ended when the conversational floor was passed from one speaker to another.

2.2. Outcome metrics

Determining the quality of an interviewee response is challenging. The literature points us to completeness and truthfulness [11], and clear articulation [12]. Politicians, however, have a reputation for equivocation and evasiveness, answering less than half the questions put to them [13]. Uncooperative politicians have been shown to make superfluous comments [14], or rely on repetition of key phrases as a diversionary tactic [15]. In this work, we regard an interview successful if the interviewee answered questions fully, directly and clearly. Accordingly, we have constructed four measures of interviewee responses: specificity, diversity, relevance, and clarity. These are broadly influenced by Gricean conversational maxims of quality, quantity, relation and manner [16].

Interviewee speech turns were tokenised, part-of-speech tagged and lemmatised using the `Stanza` neural pipeline [17]. Post calculation, all success measures were normalised.

Our measure of **specificity**, inspired by research in investigative interviewing [18, 19], quantifies interviewee references to key details such as people, objects, locations, and temporal details. We automated this process using `spaCy`'s [20] Named Entity Recognizer (NER). To account for differences in interview length, we normalise unique named entity counts over the number of noun phrases uttered by the interviewee.

Clarity measures the average concreteness of interviewee speech. Concreteness is a psycholinguistic feature that refers to the degree of ambiguity of a word. We measure word concreteness scores using an established dictionary based on prior psycholinguistic research [21]. We take the average concreteness score over all interviewee words as our clarity measure.

An unwillingness to engage in conversation can be demonstrated by self repetition. Conversely, linguistic **diversity** has been linked with honesty and trustworthiness [22] [23]. For our measure of diversity, we use the global type-token ratio of interviewee speech, i.e. we divide the number of unique interviewee words by the total number of interviewee words.

Relevance reflects the extent an interviewee's response shares semantic similarity with the question they were asked. First, each non-stop-word is transformed into a vector using pre-trained `GloVe` word embeddings [24]. We then create a turn-level vector by averaging over all word embeddings within a speech turn. We calculate the similarity of each interviewee response with the preceding question via cosine similarity (the cosine of the angle between two non-zero vectors). Relevance is computed as the mean score over all question-answer pairs.

2.3. Features

We set out to predict the success of an interview from the behaviour of the interviewer. This approach was informed by communication accommodation theory, which posits that the extent speakers converge, maintain or diverge from each other linguistically correlates with their social goals [7]. Convergence (also known as mirroring, alignment or entrainment) indicates a shared understanding between speakers [25], and is associated with success in collaborative tasks [1], and increased compliance and cooperation [26]. Furthermore, higher-levels of convergence can make an individual more likeable to those they are mirroring [27]. Conversely, reticence to converge can reflect a desire to maintain personal identity [7]. Given the dynamics of an interview, we might expect divergence to play a more prominent role in the interaction; speakers in an interview occupy particular roles, whereby the interviewer asks questions that the interviewee is expected to answer. We therefore expect the interviewer to converge on certain linguistic features, relating to the topic under discussion, and diverge on others, such as interrogatives: *who, what, where, why, when*.

2.3.1. Local accommodation

We distinguish between two types of interviewer accommodation: local accommodation and global alignment. For local accommodation (LA), we follow the probabilistic framework described in [28]. This approach computes a token-level probability of the interviewer mirroring the interviewee by comparing the probability of the interviewer speaking a word after it was spoken by the interviewee and otherwise in the conversation:

$$LA^{F}_{(i,j)} \triangleq P(T^{F}_{i}|T^{F}_{j}, T_i \hookrightarrow T_j) - P(T^{F}_{i}|T_i \hookrightarrow T_j) \quad (1)$$

Here, the term on the left, $LA^{F}_{(i,j)}$ is the local accommodation of feature F by speaker i in relation to speaker j. The first term on the right is the conditional probability of speaker i, uttering F, given its previous usage by speaker j. The second term on the right is the total probability of speaker i uttering F over all replies to speaker j. We compute a score between -1 and 1 for each feature. Positive values indicate convergence, and negative values indicate divergence.

We apply equation (1) to three separate bag-of-features models: an n-gram (NG) model, where we calculate scores for unigrams, bigrams and trigrams, restricting unigrams and bigrams to 200 features each; a lexical categorization (LC) model, based on 70 categories generated by the Linguistic Inquiry Word Count (LIWC) tool [29]; and a dependency relations (DR) model where we perform dependency parsing via `spaCy` [20], and use the dependency relation tags as features. Table 1 provides an example of pre-processed text for each model.

2.3.2. Global alignment & meta features

We constructed global alignment features based on existing computational linguistic research:

Global Language Style Matching (gLSM) is a speaker-independent similarity measure of linguistic style, i.e., speakers' alignment of semantically neutral function words such as pronouns, adjectives, and adverbs [30]. We calculate the gLSM score between the two speakers as outlined in [3], and use the score as a single feature.

For the four measures defined below, we use the mean, min, and max scores as features:

Reciprocal Language Style Matching (rLSM) measures the extent one speaker is matching the linguistic style of another on a turn-by-turn level. We calculate rLSM per interviewer speech turn as defined in [31].

Turn length difference is measured in words as the difference between each interviewer turn and the preceding interviewee turn.

Semantic relatedness measures semantic similarity via cosine similarity between each interviewer turn and the preceding interviewee turn.

Branching factor difference. In a tree-structure, the branching factor is the average number of child nodes. By performing dependency parsing on each speech turn, the branching factor becomes a proxy for syntactic depth. We compute the average branching factor over all sentences in a speech turn, and calculate the absolute difference between each interviewer turn and the preceding interviewee turn.

Variables from the meta-data collected for each interview were also included as one-hot encoded categorical variables. These included: interview length (in words); host network; political orientation of each speaker; if the speakers shared a political orientation; if the speakers were of the same gender.

2.4. Supervised machine learning

To predict each outcome metric from the features described in section 2.3, we used a forty-tree random forest with default hyper-parameters. Alternative ensemble-based supervised learning algorithms including extra trees, gradient boosting, and XGBoost achieved equal performance. Models were cross validated using K-fold, with k=10. We evaluated model performance using the root mean squared error (RMSE). We employ two baselines: an estimator that repeated the training mean (B1), and a bag-of-words model based on interviewer word frequency counts (B2).

2.5. Manual validation of computational outcome metrics

We conducted a validation study on the computational outcome metrics described in section 2.2. Ten interviews were randomly selected from the corpus, ensuring a varied distribution of outcome scores. We recruited eight human raters, naive to the purposes of the study, to assess each interview. Raters scored each interview on a one-to-ten scale per outcome metric. For example, *to what extent did the interviewer express themselves in a clear manner?* Participants also provided an overall quality score based on the perceived quality of interviewee responses. We performed an intraclass correlation (ICC) analysis on each outcome metric to measure the degree of agreement amongst our human raters. We consider ICC scores above 0.8 to indicate 'good' inter-rater agreement. The normalised computational scores were then compared to normalised human ratings.

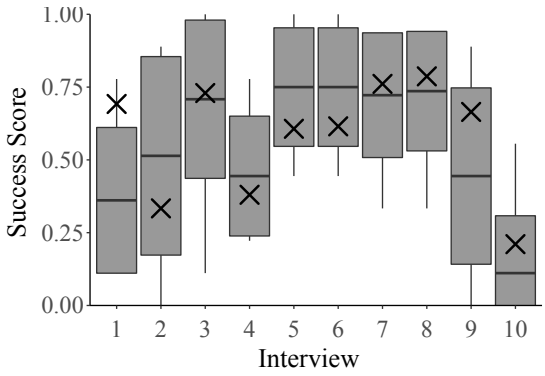

Figure 1: *Normalised distribution of human ratings of overall interview quality. The overlaid X shows the corresponding normalised computational score.*

Table 2: *Mean ($\pm SD$) RMSE scores per model iteration when predicting Clarity, Diversity, and Relevance.*

Model	Clarity	Diversity	Relevance
B1	0.144±0.016	0.158±0.016	0.142±0.011
B2	0.138±0.017	0.123±0.012	0.129±0.014
Local (All)	0.128±0.013	0.114±0.013	0.125±0.012
NG Only	0.129±0.013	0.115±0.013	0.126±0.013
LC Only	0.134±0.018	0.123±0.007	0.133±0.015
DR Only	0.134±0.015	0.121±0.008	0.137±0.014
Global	0.133±0.016	0.118±0.015	0.120±0.010
Meta	0.160±0.016	0.081±0.008	0.153±0.008
Loc. + Glo.	0.124±0.010	0.108±0.014	0.115±0.010
All	0.124±0.011	0.075±0.007	0.115±0.009

3. Results and discussion

3.1. Analysis of outcome metrics

We observed good agreement (ICC $>$.8) per outcome amongst our eight human raters. Comparing the computational scores to the normalised distribution of human ratings, we find the following percentage of computational scores that fell within one standard deviation of the mean human score: specificity = 70%, clarity = 50%, diversity = 80%, relevance = 70%.

To create an overall quality score computationally, we used a linear regression model to measure how human-raters' overall rating was weighted by their individual ratings for clarity, diversity and relevance (Specificity was omitted as our computational models did not show improvement over the baseline (B1)). The appropriate coefficients were extracted and applied to the computational scores to create an overall success score (S):

$$S = 0.23 \times Clr + 0.39 \times Div + 0.42 \times Rel \qquad (2)$$

Figure 1 shows the normalised distribution of overall success scores given by human raters per interview, with the computational score marked with an X. We find a 90% agreement with the mean human score (computational scores within one SD). Based on the average variance between the mean human score and the computational score, it would require four human raters to improve on the computational predictions.

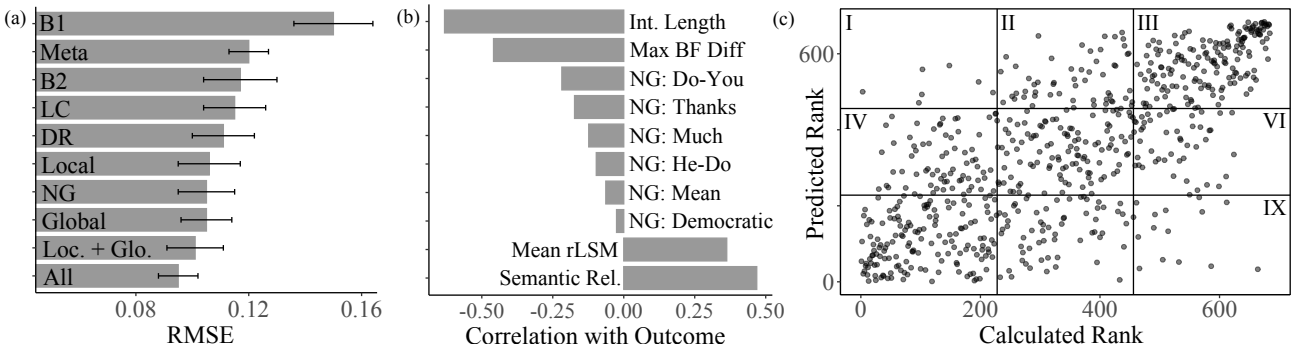

Figure 2: *2a. (left) - Prediction performance for each model iteration on the overall success measure; 2b (centre) - Correlation of the top ten features on the All Features model with the overall success score; 2c. (right) - Relationship between rank-ordered ground-truth and prediction for the Loc. + Glo. model predicting overall interview success.*

3.2. Prediction performance

Table 2 reports the RMSE for each model iteration for the clarity, diversity, and relevance metrics. The performance for overall success is shown in Figure 2 (a). One-tailed Wilcoxon signed-rank tests indicated that absolute model errors for clarity, diversity, relevance, and overall success were significantly lower than B1 and B2 ($p < .001$), but specificity was not. The combination of local and global accommodation features (Loc. + Glo. model) exceeded both baselines for clarity, diversity, relevance, and overall success. Meta features alone only exceeded the baseline when predicting diversity. The addition of the meta features to the local and global features (All model) only improved performance further for diversity and consequently for the overall success. When predicting overall success, the biggest improvement in RMSE is for the combined model (All), with a 37% improvement on B1 and 19% on B2.

In Figure 2 (b), we report the most important features for the *All* model when predicting overall success. Using permutation feature importance [32], we report the strength and direction of correlation between each feature and the outcome. Notably, *Mean rLSM* and *Semantic Relatedness* positively correlated with outcome, and *Max. Branching Factor Difference* correlated negatively. This is consistent with communication accommodation theory and shows that closer linguistic distances correlated with optimal outcomes.

The Glo. + Loc. model predicts interview success based on the interviewer's accommodation alone. To illustrate the affect of this accommodation on the success of an interview, in Figure 2 (c) we examine the relationship between predictions based on the Glo. + Loc. model and computational ground-truth when ranking the entire corpus by overall success. We can use this form of analysis to identify interviews where high levels of accommodation aligns with interview success (III), and where it does not (I). We can also spot interviews with a high success score despite a low level of measured accommodation (IX). We envision this inquiry helpful when a large corpus requires filtering before further analysis.

4. Conclusions

This paper introduces an automated cross-disciplinary approach for analysing interviews and successfully demonstrates it on a corpus of publicly available political interviews.

Our results confirm that we can successfully encode social scientific knowledge pertinent to interviewing into a computational analysis. Prudently, this can be harnessed both as a full analysis or as an initial mapping of a large corpus of conversational transcripts. Our method offers an interpretable and reproducible alternative to pre-labelled interview transcripts. This should encourage both computer scientists and social scientists alike when seeking to analyse conversations at scale.

Our psychologically-informed models significantly outperform a simple bag-of-words model, justifying domain-knowledge inclusion within computer science research. Using decipherable features also renders the analysis useful for future research within other domains. We have modelled an array of linguistic features, however, we note our choice of features is not exhaustive. Non-linguistic and paralinguistic behaviours may also contribute to accommodation.

Despite the close alignment between human and computer scores for specificity, our models did not successfully predict this measure based on the interviewer's behavior. This result may be specific to the political interview domain as establishing specific information is an unlikely goal within political interviewing. We choose to include specificity in this work as it may be of use for analysing interviews where the objective is more explicitly focused on information-gathering, for example, within the context of criminal investigation.

Our results indicate that political interviewing is a worthwhile setting to explore accommodation. A key advantage of our approach, however, is its transferability to other domains. We hope this work will lead to further adoption of the suggested cross-disciplinary approach to analyzing conversation at scale.

5. Acknowledgements

The authors would like to thank the participants who performed the manual validation task. We also thank the anonymous reviewers for providing valuable feedback. Darren Cook acknowledges support from the EPSRC and ESRC Centre for Doctoral Training in the Management and Quantification of Risk and Uncertainty in Complex Systems and Environments under ref. C110463K.

6. References

[1] D. Reitter and J. D. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, 2014.

[2] I. Naim, M. I. Tanveer, D. Gildea, Mohammed, and Hoque, "Automated Analysis and Prediction of Job Interview Performance,"

IEEE Transactions of Affective Computing, vol. 9, no. 2, pp. 191–204, 2018.

[3] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, no. 1, pp. 39–44, 2011.

[4] P. J. Taylor and S. Thomas, "Linguistic Style Matching and Negotiation Outcome," *Negotiation and Conflict Management Research*, vol. 1, no. 3, pp. 263–281, 2008.

[5] B. H. Richardson, P. J. Taylor, B. Snook, S. M. Conchie, and C. Bennell, "Language style matching and police interrogation outcomes," *Law and Human Behavior*, vol. 38, no. 4, pp. 357–366, 2014.

[6] H. H. Clark, *Using language*. Cambridge University Press, 1996.

[7] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context, and consequence," in *Contexts of Accommodation*, H. Giles, N. Coupland, and J. Coupland, Eds. New York, NY: Cambridge University Press, 1991, ch. 1, pp. 1–68.

[8] A. Schweitzer, N. Lewandowski, and D. Duran, "Social attractiveness in dialogs," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, F. Lacerda, D. House, M. Heldner, J. Gustafson, S. Strombergsson, and M. Wlodarczak, Eds. Stockholm, Sweden: International Speech Communication Association, 2017, pp. 2243–2247.

[9] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Columbus, OH: Association for Computational Linguistics, 2008, pp. 169–172.

[10] N. D. Duran, A. Paxton, and R. Fusaroli, "ALIGN: Analyzing Linguistic Interactions With Generalizable techNiques—A Python Library," *Psychological Methods*, vol. 24, no. 4, pp. 419–438, 2019.

[11] J. Rendle-Short, "Neutralism and adversarial challenges in the political news interview," *Discourse & Communication*, vol. 1, no. 4, pp. 387–406, 2007.

[12] J. Heritage, "Analyzing News Interviews: Aspects of the Production of Talk for an Overhearing Audience," in *Handbook of Discourse Analysis*, T. A. van Dijk, Ed. London, UK: Academic Press London, 1985, ch. 8, pp. 95–117.

[13] M. Waddle and P. Bull, "You're Important, Jeremy, but not that Important": Personalized Responses and Equivocation in Political Interviews," *Journal of Social and Political Psychology*, vol. 8, no. 2, pp. 560–581, 2020.

[14] B. Pluss, "Non-Cooperation in Dialogue," in *Proceedings of the ACL 2010 Student Research Workshop*, S. Demir, J. Raab, N. Reiter, M. Lopatkova, and T. Strzalkowski, Eds. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 1–6. [Online]. Available: https://www.aclweb.org/anthology/P10-3000

[15] S. Young, "The broadcast political interview and strategies used by politicians: How the Australian prime minister promoted the Iraq War," *Media, Culture and Society*, vol. 30, no. 5, pp. 623–640, 2008.

[16] P. Grice, *Studies in the Way of Words*. Harvard University Press, 1989.

[17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza : A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, A. Celikyilmaz and T.-H. Wen, Eds. Seattle, WA: Association for Computational Linguistics, 2020, pp. 101–108.

[18] K. Collins and N. Carthy, "No rapport, no comment: The relationship between rapport and communication during investigative interviews with suspects," *Journal of Investigative Psychology and Offender Profiling*, vol. 16, no. 1, pp. 18–31, 2018.

[19] L. J. Alison, E. E. Alison, G. Noone, S. Elntib, and P. Christiansen, "Why tough tactics fail and rapport gets results: Observing rapport-based interpersonal techniques (ORBIT) to generate useful information from terrorists," *Psychology, Public Policy, and Law*, vol. 19, no. 4, pp. 411–431, 2013.

[20] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[21] M. Wilson, "MRC Psycholinguistic Database : Machine Usable Dictionary," *Behavior Research Methods, Instruments, & Computers*, vol. 20, no. 1, pp. 6–10, 1988.

[22] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication," *Group Decision and Negotiation*, vol. 13, no. 1, pp. 81–106, 2004.

[23] R. Hou, V. Pérez-Rosas, S. Loeb, and R. Mihalcea, "Towards automatic detection of misinformation in online medical videos," in *2019 International conference on multimodal interaction*, 2019, pp. 235–243.

[24] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, 2014.

[25] S. Garrod and M. J. Pickering, "Why is conversation so easy?" *Trends in Cognitive Sciences*, vol. 8, no. 1, pp. 8–11, 2004.

[26] W. Kulesza, D. Dolinski, A. Huisman, and R. Majewski, "The Echo Effect: The Power of Verbal Mimicry to Influence Prosocial Behavior," *Journal of Language and Social Psychology*, vol. 33, no. 2, pp. 183–201, 2014.

[27] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[28] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark My Words! Linguistic Style Accommodation in Social Media," in *Proceedings of the 20th International Conference on World Wide Web*, S. Sadagopan, K. Ramamritham, A. Kumar, and M. Ravindra, Eds. Hyderabad, India: Association for Computing Machinery, 2011, pp. 745–754.

[29] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, *The Development and Psychometric Properties of LIWC2015*. Austin, TX: Pennebaker Conglomerates, 2015.

[30] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[31] L. C. Müller-Frommeyer, N. A. Frommeyer, and S. Kauffeld, "Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching," *Behavior Research Methods*, vol. 51, no. 3, pp. 1343–1359, 2019.

[32] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

A. Validation study

Here we describe in detail the validation study we performed and is briefly described in section 2.5 of the paper. The purpose of this small study was to measure how well our computational outcome metrics (specificity, clarity, diversity, relevance) correlated with analogous interviewee behaviours, as reported by human raters.

The ten interviews chosen for this work were selected as following: first, we initially filtered our corpus to only include video format (YouTube). This meant that our human-raters would be able to utilize paralinguistic and visual information in their evaluations in addition to the speakers' verbal behaviour. Filtering created a subset of 137 eligible interviews (approximately 20% of the entire corpus). From this, we grouped interviews into categories based on the host network (no eligible videos were identified for ABC or MSNBC). We then randomly selected three interviews from CNN, three from Fox, two from NBC, and two from CBS. These selections were spot-checked to ensure we had captured a range of speaker demographics, and interviews of varying length. See Table A.1 for information on each of the 10 interviews included in the validation study, including a link to the video.

Our eight human-raters were known to the first and second authors and volunteered to take part in the survey, dedicating roughly two hours to watch and rate the videos. All raters were adults $(18+)$, and spoke English as a first or second language. Raters were not paid or reimbursed for their efforts.

An accompanying survey was devised to capture raters evaluations of each interview. This was conducted online via SurveyMonkey, under a free basic plan (see `https://www.surveymonkey.com/`). For each interview, raters were required to answer five questions. Four of these questions were targeted at each outcome metric, with the final question an overall assessment of the responses of the interviewee. Answers could be given on an ascending $1-10$ rating scale, where higher scores indicated a more positive evaluation. Table A.2 details the questions asked.

All human raters completed the survey, for all 10 interviews, generating 80 observations. All rates watched the videos in the same order. Some rates watched all the videos in a single sitting and some over the course of several days. See Table A.3 for the anonymised raw data collected.

Figure A.1 reports the normalised distribution of ratings across the ten interviews for each outcome metric, overlaid by the corresponding computational score. By comparing the interviews to each other, we can clearly identify those interviews the human-raters felt were better or worse. For example, the tenth interview acquired the lowest scores for each of the four metrics. Inspection of the video file of this interview revealed that it was characterised by high levels of argumentation and interruption. We note that the agreement between the overall computational score and the the overall human rating of the interviews (As shown in Figure 1 in the paper) was better then the agreement between the computational and human scores of any of the individual metrics.

B. Comparison of tree-based ensemble learning algorithms

Here we describe the the alternative models tested in this work.

Our task was characterised as a series of supervised regressions, where each outcome measure would be used as the target feature. Our predictors, described in Section 2.3 of the paper, were collated from a large number of features based on bag-of-words models (see Table B.1). We tested the performance of four popular decision tree supervised learning algorithms: random forest, extra trees, gradient boosting, and XGBoost. We used the python library `Scikit-Learn` to implement each algorithm, using the default hyper-parameters (with the exception of the number of estimators, which was set at forty), and validated the models' performance via K-fold cross validation, with k=10.

Tables B.2, B.3, B.3 and B.4 report the full set of RMSE means and standard deviations when features were trained using a random forest, extra trees, gradient boosting and XGBoost, respectively.

To compare model performance we used the best performing model for each algorithm when predicting the overall success score. In each case, this was the All features model. Differences in mean RMSE between the algorithms was slight, with random forest, gradient boosting, and XGBoost all generating an average RMSE=0.095, with extra trees slightly lower at 0.094. To test for any statistical differences between the algorithms, we performed a one-way analysis of variance (ANOVA) on the RMSE scores for each fold. The ANOVA was performed having first confirmed the absence of outliers. The assumption of normality was satisfied via a non-significant Shapiro-Wilk's test ($p>$ .05). Similarly, equality of the variance of differences between algorithms was assumed via a non-significant Mauchly's Test of Sphericity ($p>$ .05). The output of the ANOVA revealed no statistically significant effect between the four algorithms: $F(3, 27) = 0.273, p > .05$. The models were therefore comparable, and the algorithm reported in the paper is the random forest.

C. Specificity

Of the four outcome metrics we explored, only our measure of specificity failed to exceed the baseline (see Table B.2). In this section we highlight our motivation behind specificity as an outcome in a political interview, and offer insight into why we did not observe a reduction in model error when specificity was the outcome measure.

Our measure of specificity was broadly influenced by work in police interrogation methods [19]. Here, the success of an interrogation is determined by, amongst other factors, the extent to which the suspect or witness introduces previously unknown information (also known as information yield). The yield of the suspect's speech is assessed over a sliding fifteen-minute window, and scored via a Likert-type scale on the degree of specific information provided. Specific information in this instance pertains to references to people, places, times, and dates, and references to motive and criminal opportunity.

The time taken to manually identify and label this information motivated us to develop an automated solution. Named Entity Recognition (NER) was chosen as names, places, times, and dates are types of named entity, and are commonly sought features in information retrieval tasks. We used `spaCy's` pretrained NER to identify named entities in interviewee speech turns. Of the eighteen entity types recognised, we selected twelve we felt were relevant in a political context (see Table C.1). We ignored any repetition of named entities, instead taking the count of unique references only. To create a normalised score, we divided the count of unique named entities by the number of noun phrases (a sequence of words where the head word is a noun) produced by the interviewee. We found that our computational measure performed comparably with

Table A.1: *Demographic information of the ten interviews used in the validation study, and links to video files*

Network	Interviewer	Interviewee	Length (mins:secs)	Link (https://www.youtube.com/)
Fox	Chris Wallace	Stephen Miller	13 : 25	/watch?v=vXUWHk7sqe0
CBS	Gayle King	Ivanka Trump	7 : 03	/watch?v=XRLdnBpEMAA
NBC	Chuck Todd	Beto O'Rourke	7 : 53	/watch?v=1ZG6pku_pWY
CNN	Jake Tapper	Rudy Giuliani	15 : 21	/watch?v=Tn7MsHcamdU
CBS	John Dickerson	Marco Rubio	9 : 17	/watch?v=RsY06MOgXj0
CNN	Wolf Blitzer	Eric Swalwell	9 : 08	/watch?v=zf7ygAN2vf4
Fox	Tucker Carlson	Tulsi Gabbard	5 : 56	/watch?v=ZuUaAyYzBwc
NBC	Chuck Todd	Hakeem Jeffries	8 : 11	/watch?v=v1d1PU9nr24
Fox	Chris Wallace	Val Demings	10 : 41	/watch?v=leNmghMf5wI
CNN	Anderson Cooper	Kellyann Conway	25 : 53	/watch?v=-otxWE6dBxk

Table A.2: *Manual validation survey questions, high responses, and low responses per outcome measure*

Specificity	Question	How informative were the interviewee's responses to the interviewer's questions?
	High Response	Provided a lot of specific information
	Low Response	Did not provide any factual information irrespective of the question
Clarity	Question	To what extent did the interviewee express themselves in a clear manner?
	High Response	Straightforward, easy to understand, only one interpretation is possible
	Low Response	Vague, ambiguous, impossible to understand, bears no relevance to actual events
Relevance	Question	How relevant were the interviewee's responses to the interviewer's questions?
	High Response	Interviewee stayed on track and addressed the question
	Low Response	Interviewee typically did not address the question
Diversity	Question	Were the interviewee's answers repetitive or diverse?
	High Response	Interviewee used a broad range of words and phrases
	Low Response	Interviewee repeated a small number of words or phrases throughout
Overall	Question	Considering your previous answers, how would you rate the overall quality of the interviewee's responses in this interview?
	High Response	Very high quality, easy to follow, with a consistently high level of information provided
	Low Response	Very poor quality, difficult to follow, with little relevant or specific information provided

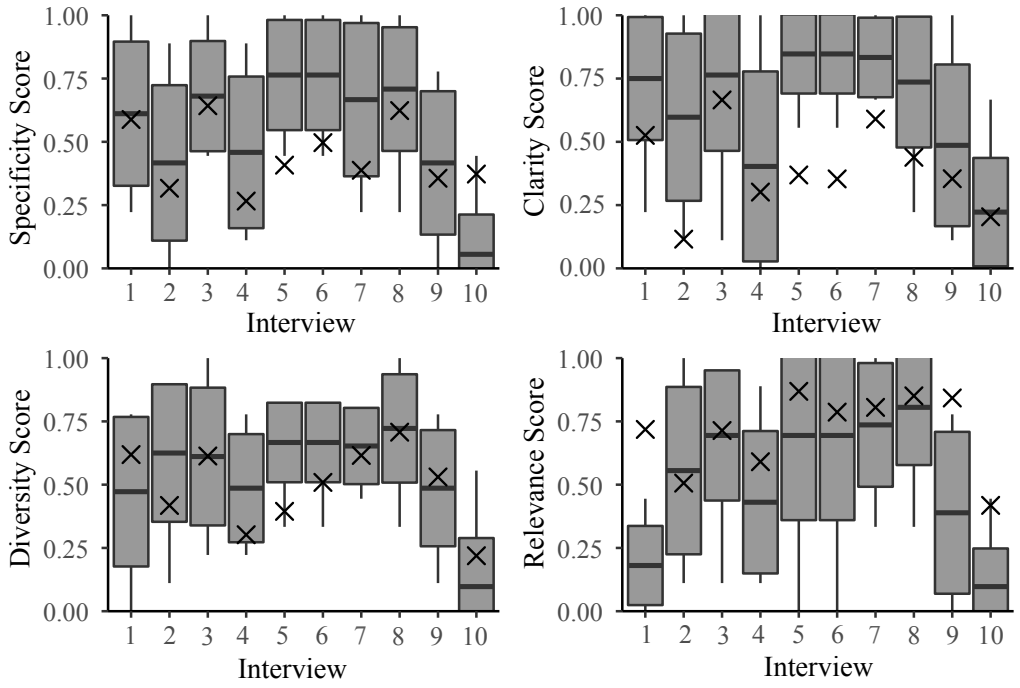

Figure A.1: *Normalised distribution of human ratings for specificity (top left), clarity (top right), diversity (bottom left), and relevance (bottom right). X refers to the normalised computational score predicted per interview, using all features*

Table A.3: *Raw scores provided by each rater per interview on the manual validation task*

Interview	Participant	Specificity	Clarity	Diversity	Relevance	Overall
1	1	9	2	10	8	8
	1	9	2	10	8	8
	2	6	2	3	3	2
	3	8	4	9	6	5
	4	3	3	7	4	3
	5	7	1	7	1	2
	6	6	5	9	8	7
	7	3	3	9	4	3
	8	10	1	8	8	4
2	1	3	8	9	5	6
	2	1	2	2	2	2
	3	4	4	4	5	4
	4	7	10	10	9	9
	5	7	7	8	8	8
	6	9	8	7	7	9
	7	5	7	8	9	6
	8	2	2	3	8	1
3	1	9	8	9	6	8
	2	5	2	2	3	2
	3	5	8	10	5	8
	4	8	8	8	8	8
	5	8	8	7	9	8
	6	7	9	10	7	9
	7	10	9	10	10	10
	8	5	6	7	4	6
4	1	6	7	5	6	6
	2	4	4	3	3	3
	3	3	6	3	5	5
	4	3	2	5	3	4
	5	9	9	9	7	8
	6	9	2	1	7	4
	7	2	3	1	4	3
	8	5	6	10	8	7
5	1	8	10	10	6	8
	2	10	10	10	8	10
	3	8	9	8	8	9
	4	8	7	8	7	8
	5	9	8	9	8	8
	6	10	8	10	8	9
	7	5	5	8	4	5
	8	5	1	6	7	5
6	1	8	10	10	6	8
	2	10	10	10	8	10
	3	8	9	8	8	9
	4	8	7	8	7	8
	5	9	8	9	8	8
	6	10	8	10	8	9
	7	5	5	8	4	5
	8	5	1	6	7	5
7	1	8	10	10	8	9
	2	9	10	10	7	9
	3	3	5	7	5	6
	4	9	8	8	8	9
	5	10	8	10	8	9
	6	3	7	7	5	4
	7	6	4	7	6	6
	8	8	9	9	8	8
8	1	9	10	9	7	9
	2	7	10	9	8	8
	3	8	10	9	7	9
	4	9	9	9	10	9
	5	10	8	9	10	9
	6	7	8	8	7	7
	7	6	7	5	7	6
	8	3	4	3	4	4
9	1	5	7	6	5	5
	2	6	8	10	8	9
	3	1	1	2	2	1
	4	8	8	9	7	8
	5	8	2	3	5	4
	6	4	5	5	7	6
	7	4	3	5	6	5
	8	2	2	3	3	2
10	1	1	1	3	1	1
	2	1	2	3	1	1
	3	1	2	4	2	1
	4	5	5	7	6	6
	5	1	1	2	1	1
	6	1	2	1	1	3
	7	1	1	1	1	1
	8	1	1	3	2	2

Table B.1: *Number of features per model*

Model	Features
Local (All)	640
NG Only	527
LC Only	70
DR Only	43
Global	13
Meta	10
Loc. + Glo.	653
All	663

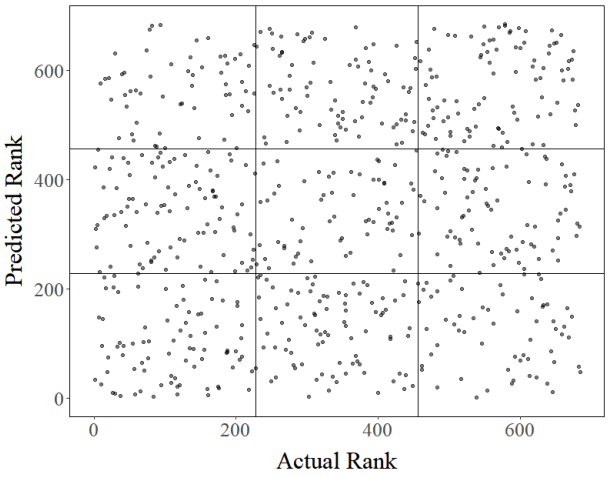

Figure C.1: *Relationship between rank-ordered actual and predicted values for specificity using all features*

how a human-being would evaluate specificity (Figure A.1), with the computational score falling within one standard deviation of the mean human rating in 70% of cases.

Figure C.1 shows that our specificity measure was not well predicted by our model. This may be application specific, as specificity is not a clear objective for the majority of political interviewers. This is further validated when fitting a linear regression to the normalised human ratings. Removing specificity did not have a major impact on the fit of the linear regression, reducing the R^2 from 0.92 to 0.91.

We decided to include the specificity measure in the paper as it may be of use to similar applications under a different domain, in particular, criminal justice.

D. Feature analysis for diversity, relevance & clarity

Figure D.1 illustrate the prominent features for the diversity, relevance, and clarity metrics. Interestingly, we observed that the maximum branching factor difference negatively correlated with both clarity and relevance. This indicates that interviewers who matched the syntactic depth of the interviewee generated clearer, and more relevant responses. Similarly, average semantic relatedness positively correlated with clarity and relevance, meaning that interviewers who consistently followed the semantic quality of the interviewee also led to better outcomes. These findings are consistent with the view that shortened linguistic distances align with collaborative social outcomes.

Curiously, while the maximum branching factor difference negatively correlated with relevance, minimum branching factor difference positively correlated. This suggests that becoming too similar, at least on a syntactic level, had a detrimental impact on how relevant the interviewees' responses were. Potentially, the minimum branching factor is capturing a linguistic distinction between the two roles, the interviewer is asking questions, and the interviewee is answering. Our models also utilise phrases that are highly-specific to the interviewer, such as the bigrams *do you*, and *you say*. Perhaps unsurprisingly, these features correlate negatively with our outcome measures. This suggests that divergence of these phrases by the interviewer aligned with clearer, and more diverse responses from the interviewee.

The most prominent feature over all the models was interview length in relation to the diversity measure. This relationship is a known consequence of using the type-token ratio (TTR) to measure language diversity. Additionally, the effect is likely being amplified by the wide range of interview lengths within our corpus. Alternative approaches that could be explored in future, such as content diversity or redundancy, are detailed in [22].

Figure D.2 compares the predicted and computationally calculated rank for all interviews for clarity, diversity and relevance. Diversity presents the most linear shape, although this is most likely due to its strong correlation with interview length. Both relevance and clarity appear less linear then the overall success measure shown in Figure C.1.

E. Error analysis

We performed an exploratory error analysis on the model (all features), plotting the absolute errors against each of the top ten important features and the score itself. We observed an increased model error at the extremes of the distribution of overall success scores (see Figure E.1), but no other pattern emerged. Specifically, we note a concentration of higher error at the top end of the distribution. This is a result of the random forest not being able to extrapolate outside of the training set, and the sparsity of data-points at the tail ends of the distribution.

F. Hyper-parameter tuning

Hyper-parameter tuning was performed on the best performing model when predicting the overall success measure (All features model). We used nested cross-validation (CV) to identify the optimal settings of the random forest. To prevent data leakage between training and testing sets, we performed two sets of *K*-fold loops, often referred to as an outer and inner loop. The outer loop operates in an identical manner to a standard *K*-fold, creating *K* distinct folds from the data. Then, within each outer fold, we perform another, typically smaller *K*-fold to optimise parameters. In this work, we maintained *K*=10 to evaluate the model, and set *K*=3 to define hyper-parameters.

We used the `GridSearchCV` package from `Scikit-Learn` to test a range of possible values for six random forest hyper-parameters. The search range for each hyper-parameter (see Table F.1).

We found that the default random forest performed equivalently well as the optimised model (default: M= 0.095, SD=0.007; optimised: M= 0.094, SD=0.012), indicating we had reached a point of diminishing returns. Results of a paired-samples *t*-test confirmed that there was no statistical difference between the default and optimised models.

Table B.2: *Mean ($\pm SD$) RMSE scores per model iteration for Random Forest*

Model	Specificity	Clarity	Diversity	Relevance	Overall
B1	0.157±0.012	0.144±0.016	0.158±0.016	0.142±0.011	0.15±0.014
B2	0.153±0.012	0.138±0.017	0.123±0.012	0.129±0.014	0.117±0.013
Local (All)	0.157±0.012	0.128±0.013	0.114±0.013	0.125±0.012	0.106±0.011
NG Only	0.158±0.013	0.129±0.013	0.115±0.013	0.126±0.013	0.105±0.01
LC Only	0.16±0.01	0.134±0.018	0.123±0.007	0.133±0.015	0.115±0.011
DR Only	0.158±0.01	0.134±0.015	0.121±0.008	0.137±0.014	0.111±0.011
Global	0.162±0.012	0.133±0.016	0.118±0.015	0.120±0.010	0.105±0.009
Meta	0.173±0.013	0.160±0.016	0.081±0.008	0.153±0.008	0.12±0.007
Loc. + Glo.	0.156±0.012	0.124±0.010	0.108±0.014	0.115±0.010	0.101±0.01
All	0.156±0.012	0.124±0.011	0.075±0.007	0.115±0.009	0.095±0.007

Table B.3: *Mean ($\pm SD$) RMSE scores per model iteration for Extra Trees*

Model	Specificity	Clarity	Diversity	Relevance	Overall
B1	0.157±0.012	0.144±0.016	0.158±0.016	0.142±0.011	0.15±0.014
B2	0.162±0.012	0.138±0.018	0.124±0.012	0.133±0.014	0.118±0.013
Local (All)	0.158±0.013	0.13±0.012	0.111±0.014	0.126±0.013	0.101±0.012
NG Only	0.157±0.013	0.128±0.011	0.114±0.013	0.127±0.014	0.102±0.012
LC Only	0.161±0.01	0.132±0.017	0.124±0.009	0.132±0.014	0.117±0.01
DR Only	0.157±0.009	0.135±0.015	0.119±0.007	0.132±0.014	0.112±0.01
Global	0.162±0.014	0.133±0.013	0.117±0.015	0.117±0.008	0.103±0.01
Meta	0.197±0.014	0.175±0.011	0.089±0.008	0.174±0.013	0.132±0.008
Loc. + Glo.	0.157±0.015	0.123±0.01	0.106±0.013	0.116±0.01	0.1±0.011
All	0.157±0.012	0.124±0.009	0.075±0.007	0.115±0.01	0.094±0.007

Table B.4: *Mean ($\pm SD$) RMSE scores per model iteration for Gradient Boosting*

Model	Specificity	Clarity	Diversity	Relevance	Overall
B1	0.157±0.012	0.144±0.016	0.158±0.016	0.142±0.011	0.15±0.014
B2	0.153±0.012	0.138±0.017	0.123±0.012	0.129±0.014	0.117±0.013
Local (All)	0.158±0.012	0.135±0.013	0.111±0.012	0.127±0.01	0.104±0.01
NG Only	0.156±0.013	0.133±0.013	0.112±0.013	0.125±0.012	0.106±0.01
LC Only	0.16±0.009	0.133±0.018	0.123±0.006	0.133±0.013	0.114±0.009
DR Only	0.157±0.011	0.132±0.015	0.123±0.007	0.131±0.015	0.11±0.011
Global	0.161±0.013	0.131±0.014	0.118±0.014	0.116±0.011	0.104±0.009
Meta	0.159±0.01	0.143±0.016	0.075±0.007	0.138±0.012	0.109±0.008
Loc. + Glo.	0.156±0.014	0.126±0.011	0.104±0.013	0.118±0.011	0.099±0.01
All	0.157±0.014	0.127±0.01	0.073±0.01	0.117±0.012	0.095±0.006

Table B.5: *Mean ($\pm SD$) RMSE scores per model iteration for XGBoost*

Model	Specificity	Clarity	Diversity	Relevance	Overall
B1	0.157±0.012	0.144±0.016	0.158±0.016	0.142±0.011	0.15±0.014
B2	0.156±0.011	0.135±0.018	0.123±0.011	0.133±0.01	0.116±0.012
Local (All)	0.161±0.012	0.132±0.011	0.11±0.012	0.124±0.013	0.103±0.008
NG Only	0.16±0.013	0.132±0.012	0.115±0.011	0.126±0.012	0.103±0.008
LC Only	0.16±0.01	0.136±0.017	0.12±0.008	0.135±0.014	0.114±0.013
DR Only	0.161±0.01	0.133±0.014	0.121±0.006	0.129±0.015	0.108±0.011
Global	0.166±0.015	0.138±0.014	0.12±0.012	0.12±0.011	0.106±0.009
Meta	0.164±0.013	0.149±0.016	0.077±0.008	0.143±0.011	0.113±0.007
Loc. + Glo.	0.159±0.013	0.124±0.009	0.104±0.013	0.117±0.011	0.101±0.008
All	0.16±0.013	0.124±0.009	0.076±0.009	0.117±0.012	0.095±0.008

Table C.1: *Named Entity classes used to measure specificity*

Entity Type	Class Label
Person	PERSON
Nations, Religions, Political Groups	NORP
Infrastructure	FAC
Organizations	ORG
Countries, States	GPE
Non-GPE Locations	LOC
Battles, wars	EVENT
Named laws	LAW
Absolute dates	DATE
Times less than a day	TIME
Money	MONETARY
Measurements	QUANTITY

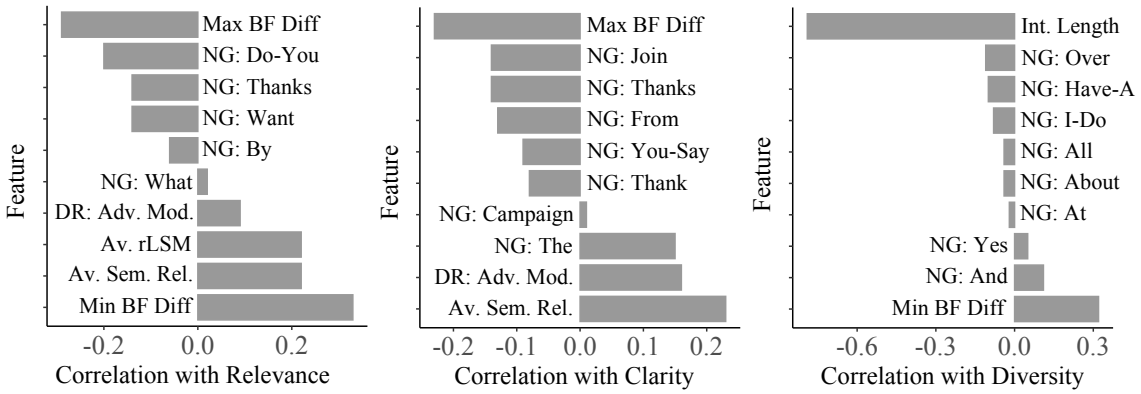

Figure D.1: *Correlation of the top ten features for predicting relevance, clarity, and diversity using the 'all features' model*

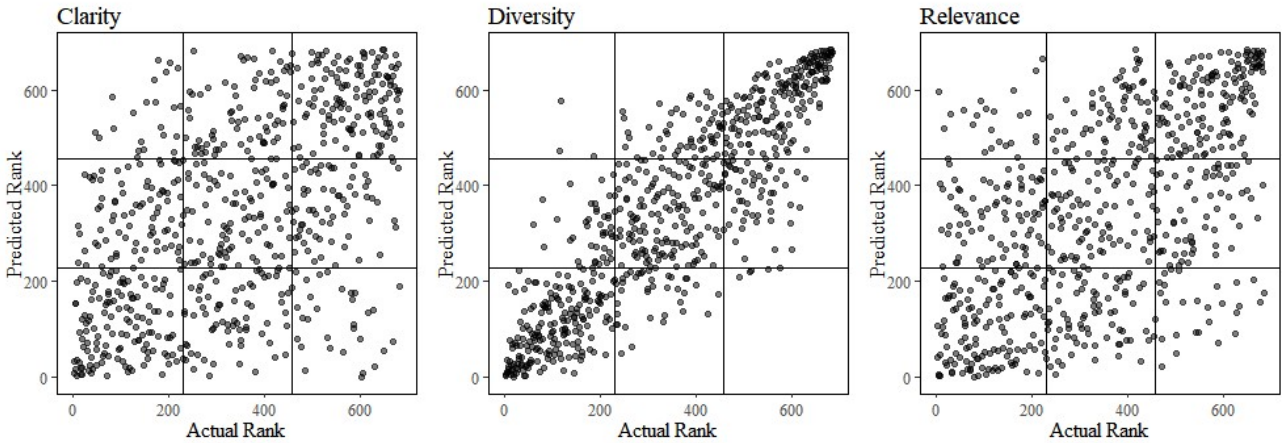

Figure D.2: *Relationship between rank-ordered actual and predicted for clarity (left), diversity (centre), and relevance (right) using all features*

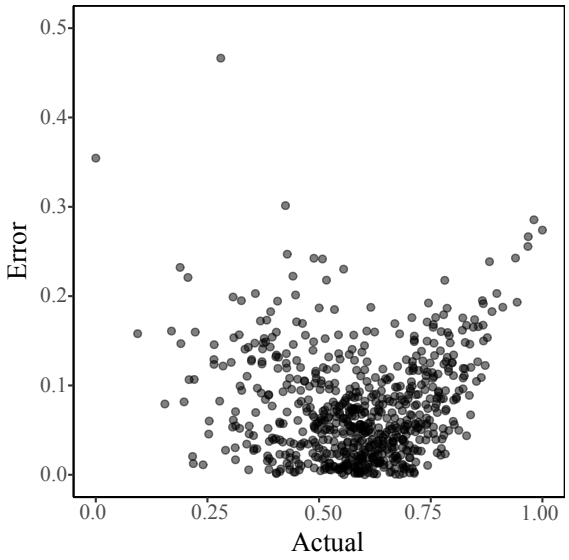

Figure E.1: *Relationship between the absolute model errors (y axis) and actual score (x axis) when predicting overall success with the all features model*

Table F.1: *Range of hyper-parameter settings used in the grid-search*

Description	Parameter	Range of Values
Number of trees	n_estimators	40, 80, 120, 160
Max. depth of trees	max_depth	No Maximum., 10, 15
Max. features to consider when splitting node	max_features	n_estimators, $\sqrt{n\_estimators}$, $\log_2 n\_estimators$
Min. samples to split a node	min_samples_split	2, 3
Min. samples at a leaf node	min_samples_leaf	1, 2, 3
Min. impurity decrease required to split node	min_impurity_decrease	0.0, 0.1

G. A note on diversity as a success measure in political interviews

Diversity was included as a success measure in an attempt to capture the degree of openness expressed by the interviewee. A reduction in lexical diversity has been linked with deceptive behaviour in speech [22]. There is also tentative support to the claim that measures of lexical diversity predict misinformation [23]. However, we acknowledge that the level of diversity does not always reflect the quality of an interviewee's verbal behaviour. There are circumstances where political communication is aided by lower rates of linguistic diversity. For example, repetition of particular phrases may be used to emphasise key messages during an election campaign. We might also expect politicians to deliberately use a smaller vocabulary in order to appeal to a greater majority of voters. However, we do not believe this detracts from our overall approach. Our model is flexible and can easily fit our accommodation features to an improved measures of conversational outcomes.

H. A note on transcriptions generated from secondary sources

The transcripts used in this study were secondary sources derived from opportunistic sampling of online repositories. We have therefore performed additional pre-processing to ensure both accuracy and consistency. First, we spot-checked the accuracy of the transcription against video footage of the interview, if available on YouTube. Whilst we could only locate video footage for approximately 20% of the corpus, we were satisfied with the general quality of the transcripts. We did, however, remove twelve interviews that had been heavily edited.

We assumed that the networks used different transcription services to generate the transcripts. We therefore performed extensive pre-processing to ensure a suitable level of inter-network consistency within the corpus. All transcripts used an orthographic method of transcription. This means they used standard spelling, and did not include any false starts or filler utterances such as 'er' or 'umm'. A proportion of the corpus did include symbols that were used to indicate hesitation or interruption. For example, a speech turn that was interrupted was often appended with a '–' sequence, and attempts to re-establish the conversational floor often prepended with the same sequence. Similarly, a hesitation was often marked with a ' - ' sequence. Whilst a possible source of accommodation in their own right, there was insufficient coverage of these non-linguistic behaviours to motivate their inclusion in this work. As such they were removed.

I. Availability of python scripts

The python scripts used in the transcript pre-processing, outcome and accommodation modelling steps described in section 2 can be found at https://github.com/cookie1986/interview-accommodation under an MIT licence.