Using Interaction Style Dimensions to Characterize Spoken Dialog Corpora

Nigel G. Ward
Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968, USA
nigelward@acm.org

Abstract

The construction of spoken dialog systems today relies heavily on appropriate corpora, but corpus selection is more an art than a science. As interaction style properties govern many aspects of dialog, they have the potential to be useful for relating and comparing corpora. This paper overviews a recently-developed model of interaction styles and shows how it can be used to identify relevant corpus differences, estimate corpus similarity, and flag likely outlier dialogs.

1 Motivation

Today the process of selecting corpora for dialog systems training or tuning is rarely systematic. This is a problem because dialog systems developers rely heavily on machine learning from corpora to acquire the various knowledge and parameters needed for effective systems. Models for predicting likely corpus suitability would therefore be very useful, but existing methods for corpus comparison rely mostly on lexical and topic overlap, e.g. (Pavlick and Nenkova, 2015), making it hard to predict how well other types of knowledge will transfer.

The scientific investigation of dialog behaviors is similarly impeded by corpus choice issues. Different research teams choose corpora to study for all sorts of reasons, leading to a healthy diversity, but also to many contradictory findings (Egger et al., 2014; Wright et al., 2019; Levitan, 2020). Methods for systematically describing corpus properties could help resolve these, potentially enabling the field of computational pragmatics to clearly describe the realm of validity of each generalization.

This paper focuses on interaction style, as this is an essential issue in providing high quality user experiences (Marge et al., 2022). This is, moreover, no longer a distant goal, as core speech components have advanced to the point where it is becoming possible to implement situation-appropriate turn taking, politeness behaviors, rapport building strategies, and so on (Metcalf et al., 2019). Because our fundamental knowledge in these areas are still spotty, developers rely on discovery or learning from corpora. Indeed, it is still common for a new development project to start with the collection of a new corpus, specific to the task, domain, user demographic, system persona and so on. Instead, we would like to be able to better exploit existing resources (Kashyap et al., 2021). One recent success was a socially well-behaved recommendation system for movies, created by discovering behaviors from a suitable subset of Switchboard data (Pecune et al., 2019). Selection of this subset was easy because Switchboard was designed around topics, and in particular the "movies" tag was available. However, we would like to be able to more precisely delineate relevant corpus subsets, and to do so even when annotations are lacking.

This paper introduces three ways to characterize spoken dialog corpora and their subsets.

2 Precursor Work

Biber, in his landmark contribution to style description, investigated what he termed "conversation text types" (Biber, 2004). Using transcripts from various corpora as data and a text-based feature set, he used Principal Component Analysis to derive three dimensions of variation, and showed how different conversations could be automatically located in this space.

This method has been very influential in the comparison of diverse text corpora, and also occasionally for speech corpora (Shen and Kikuchi, 2014). However these models generally seem to have low explanatory power; for example, Biber's three dimensions accounted for only 36% of the variance. Further, although acoustic-prosodic features potentially provide much more information than text, these have been used in corpus selection so far only by Siegert *et al.* (2018), who demonstrated their

Proceedings of the SIGdial 2022 Conference, pages 225–230
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

value, but only for the narrow problem of training emotion recognizers. Overall, work in this tradition appears not to have found practical use.

In contrast to text-based models (Troiano et al., 2021, submitted), styles in spoken dialog, and in particular interaction styles, have been less studied. Much work in this area has built on Tannen's seminal observations on "conversational styles" (Tannen, 1989, 1980). Importantly, these are not fixed properties of speakers, and frequently vary even in the course of a conversation (Dingemanse and Liesenfeld, 2022).

More recently, computational models have been developed to study style in dialog (Grothendieck et al., 2011; Laskowski, 2016; Yamamoto et al., 2020; Ward, 2021a). These works have variously used features of turn-taking and prosodic and other behaviors to derive models of style. However these models have previously been applied only to questions of how individuals vary in style, not to corpus characterization.

3 Model Properties

The explorations reported in this paper build on our own model of interaction style variation (Ward, 2021a; Ward and Avlia, 2022, submitted), because it is the most comprehensive and because the code is available. The purpose of this section is only to explain the model briefly while clarifying the aspects not clear in (Ward, 2021a) but relevant for the current exploration.

For current purposes, the model serves to take as input one or more 30-second fragments of American English conversation, and to output a representation of its style as a vector of length 8: that is, it maps dialogs into a vector space representation of interaction styles. While for current purposes this is used as a black box, it may be worth overviewing the steps of the process.

1. Low-level (frame level) prosodic features are computed, specifically the raw pitch, intensity, and cepstral coefficients.
2. These are normalized by track.
3. Filters and aggregation processes are applied to obtain mid-level features over various temporal spans, including estimates of intensity, speaking rate, phoneme lengthening, creakiness, enunciation or reduction, and the extent to which the pitch is high or low, or wide or narrow.
4. These mid-level features are normalized using parameters that brought each to mean 0 and standard deviation 1 on the training data.
5. The match of these normalized features to 12 meaningful temporal configurations is computed every 20 milliseconds. These meaningful temporal configurations represent specific American English prosodic constructions, which mark activities such as turn switch, topic closing, enthusiasm, positive assessment, empathizing, and contrasting (Ward, 2019). These cover a wide range of dialog states, activities, behaviors and interactive events.
6. The match values are binned and pooled across each 30-second fragment. There are 7 bins per configuration, thus there are bins for when a speaker is expressing a strong, mild, or weak contrast, or managing an ambiguous, clear, or strong turn switch, and so on.
7. The resulting 84 values are rotated, using Principal Component Analysis, to a representation where the top dimensions capture most of the variance.
8. The top 8 dimensions are retained. (This is because these 8 already explain 52% of the variance, because the lower dimensions lacked clear interpretations, and because including more dimensions did not significantly change the qualitative picture presented below.)

Further, each of the eight dimensions can be given an interpretation, as summarized in Table 1. Those for Dimensions 4 and 7 differ from those given by Ward (2021a), for reasons explained in (Ward and Avlia, 2022, submitted); for all, we here provide clearer descriptions. While the interpretations are not needed for most purposes, they help to understand how and whether the model is working, so the rest of this section elaborates. Evidence and further discussion appears at the companion website (Ward, 2021b).

Dimension 1 relates simply to the amount of shared engagement. Dimension 2 is very high or low when one speaker *versus* the other is taking an active speaking role and the other an active listening role. Dimension 3 involves expressing positive assessment, for example when talking about the speaker's dog, a good fishing day, or a favorite football team, *versus* expressing negative feelings,

1	13%	both participants engaged	...	lack of shared engagement
2	11%	focal speaker mostly talking	...	focal speaker listening actively
3	8%	positive assessment	...	negative feelings
4	5%	focal speaker speaks knowledgeably	...	nonfocal speaker speaks knowledgeably
5	5%	factual	...	thoughtful
6	4%	accepting things beyond individual control	...	envisioning positive change
7	3%	making points	...	referencing shared experiences
8	3%	unfussed	...	emphatic

Table 1: Inferred functions of the top 8 dimensions of interaction style. The second column shows the amount of variance explained by each dimension.

for example about underprepared students or immoral politicians. Dimension 4 is very high or low when one speaker *versus* the other is being confident and/or dominant as they talk about something they know well, while the other is acknowledging the other as an expert on the topic. For Dimension 5 the positive pole involves a thoughtful style and the negative pole a factual style, characterized, among other things, by long regions of low pitch expressing a stance of calm rationality, as the speaker describes something they know well, such as how a network is set up or how security cameras work. Dimension 6 relates to a resigned attitude, for example when taking about high rents or working in a job where there is no opportunity to meet the customers, *versus* a positive, change-oriented outlook, for example when discussing new exercise regimens, changes in women's roles, or medical research advances. Dimension 7 relates to stating and justifying opinions, for example general ideas about dealing with people or situations, *versus* finding common ground, for example when talking about similar experiences with catalog shopping, making hamburger, or drug testing. Dimension 8 involves the continuum between talk about remote or currently unimportant and half-understood or half-remembered ideas or events *versus* expressing strong opinions, for example regarding people or practices that are strongly disliked or strongly admired.

4 Use 1: Corpus Characterization

This model supports visualization of corpus differences. As an example, if we view Switchboard (Godfrey et al., 1992) as a collection of subcorpora, one per topic, we can map them out, for example by plotting the average interaction style of all fragments within that topic. Figure 1 shows this for Dimensions 1 and 3. (Projections onto other dimensions are available at the companion website.) To avoid clutter, the figure show only topics for which there was ample data (225 minutes or more) or which were among the most distinctive topics, in terms of distance on these two dimensions from the global average style. Table 2 shows the values for all 8 dimensions for the topics discussed below.

The positions of the topics in the figure suggest that the model is at least picking up something meaningful. It is informative to consider further some of the topics that appear, at first glance, to be misplaced. For example, it may seem strange that the model characterizes conversations on the topic of "metric system" as positive in style, but listening to examples shows that these conversations are mostly by engineers, who indeed discussed it positively. It may also seem strange that "woodworking" and "painting" are placed differently, as both can be at-home hobbies and projects. According to the model, their interaction styles are very different, as seen in Table 2. In particular, these suggest that dialogs about woodworking exhibited less shared engagement and were more positive and thoughtful in tone (Dimensions 1, 3, and 5, respectively). Listening confirmed that these differences were real, and likely attributable to the tendencies for woodworking to be discussed fondly by dedicated hobbyists, and painting to be discussed by novices talking about difficulties. Thus the model captures much more than simple topic similarity.

In general, diagrams like these may help researchers and developers understand the diversity within and between corpora.

5 Use 2: Similarity Estimation

This model also supports similarity estimation (Kilgarriff and Rose, 1998), for now by simply us-

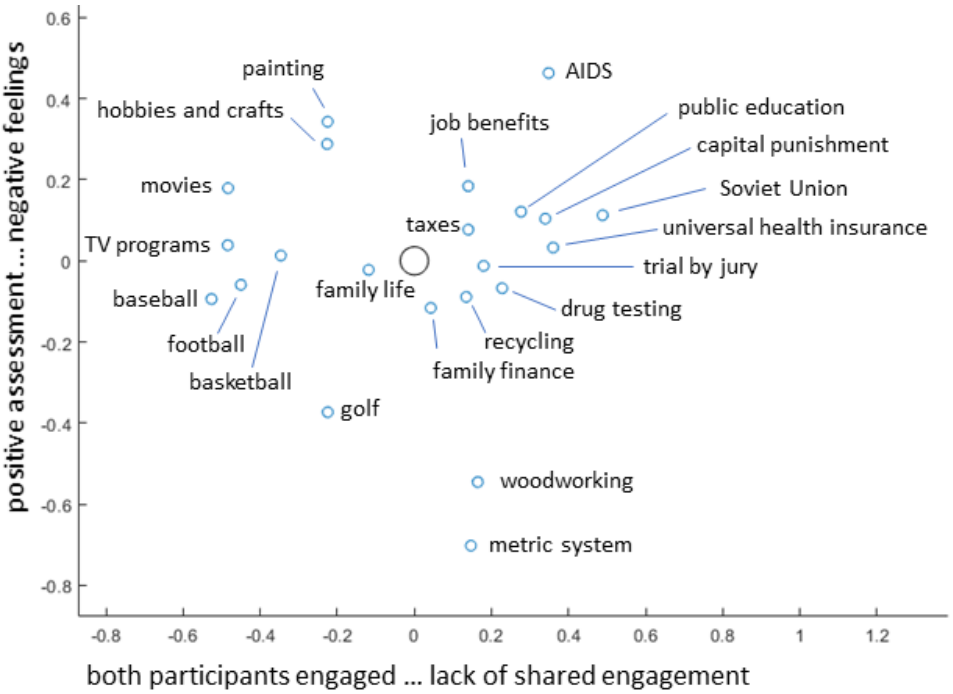

Figure 1: Average Interaction Styles of Some Topics in Switchboard, Projected to Interaction Style Dimensions 1 and 3. The large circle marks (0,0), the global average style. The axis units are standard deviations computed over all conversation fragments. The topic names shown are just mnemonics for the sentence-length prompts given to the participants.

	dimension							
	1	2	3	4	5	6	7	8
woodworking	0.6	2.2	–1.4	1.4	1.1	–0.4	0.6	0.5
painting	–0.8	3.0	0.9	1.8	–0.6	0.1	–0.6	0.5
politics	1.0	2.6	0.1	1.6	0.4	0.1	0.1	–0.2
capital punishment	1.1	2.7	0.3	1.6	0.5	0.0	0.0	–0.0
movies	–1.6	–0.0	0.5	0.0	–0.7	–0.5	0.3	–0.1

Table 2: Average interaction style for selected topics from Switchboard on the 8 dimensions.

ing the Euclidean distance in the 8-dimensional space. For example, considering Switchboard's 20 topics most distant from the global average, the closest pair was "politics" and "capital punishment," as seen in Table 2 respectively. The other most similar pairs were "baseball" and "football," "weather/climate" and "vacation spots," and "movies" and "TV programs."

Such similarity estimates could be used to support targeted data augmentation. Considering again the scenario of seeking data to train a movie recommendation system, the subcorpora closest to "movies" were "TV programs," "clothing and dress," "football," and "baseball," indicating that these would be likely be most compatible as supplementary data.

6 Use 3: Identifying Outliers

The similarity metric could also be used in support of data cleaning. For example, many conversations in Switchboard have the "movies" tag, but not all fragments are good exemplars of the typical style for talking about movies. The model can help identify these, as fragments distant from the average interaction style for this topic. For the movies topic, examination of the five most distant fragments revealed that these were indeed mostly atypical — two involved strong moral judgments, and one was mostly about audience behavior — and would be good candidates for exclusion from the training set

for a movie recommending system with a normal, upbeat style.

7 Prospects

Spoken data is fundamentally richer than text data, and recent work is exploiting this to create more informative models of corpus similarities and differences. This brief report has proposed new ways to exploit one such model, involving interaction style.

Eventually, direct quantitative evaluation of this method should be done. One way would be to examine the correspondences to human judgments of interaction styles and style similarities. This would be a long-term project, but potentially of great benefit for systematizing the scientific study of dialog phenomena.

In the short term, we think the value of these methods will instead be shown by their practical value: their ability to support the creation of better-tailored dialog systems, and to reduce the data-collection efforts required to develop them. More specifically, in addition to the three ways illustrated above, we conjecture that the model will be useful in at least three other ways. 1) It could support quality control and consistency control during corpus collection. 2) It could support attempts to collect corpora with a sweet-spot style that is simultaneously natural for humans and implementable with current technology (Budzianowski et al., 2018; Byrne et al., 2019), by identifying the dimensions in which such corpora most resemble both human-human dialogs and technically-realizable dialogs. 3) It could support the development of widely useful pretrained models by supporting the selection of truly diverse sets of dialog corpora.

To support such uses, the code is available at (Ward, 2021c).

Acknowledgments: I thank Jonathan E. Avila for helping refine the dimension interpretations.

References

Douglas Biber. 2004. Conversation text types: A multi-dimensional analysis. In *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, pages 15–34. Presses Universitaires de Louvain.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz: a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Empirical Methods in Natural Language Processing*, pages 5016 – 5012.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Empirical Methods in Natural Language Processing*, page 4515–452.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *ACL*, pages 5614–5633.

Sebastian Egger, Peter Reichl, and Katrin Schoenenberg. 2014. Quality of experience and interactivity. In Sebastian Moeller and Alexander Raake, editors, *Quality of Experience*, pages 149–161. Springer.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520.

John Grothendieck, Allen L. Gorin, and Nash M. Borges. 2011. Social correlates of turn-taking style. *Computer Speech and Language*, 25:789–801.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: a survey and empirical analysis. In *NAACL*, pages 1830–1849.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 46–52.

Kornel Laskowski. 2016. A framework for the automatic inference of stochastic turn-taking styles. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–211.

Rivka Levitan. 2020. Developing an integrated model of speech entrainment. In *IJCAI*, pages 5159 – 5163.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, et al. 2022. Spoken language interaction with robots: Research issues and recommendations. *Computer Speech and Language*, 71.

Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. 2019. Mirroring to build trust in digital assistants. In *Interspeech*, pages 4000–4004.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224.

Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 135–143.

Raymond Shen and Hideaki Kikuchi. 2014. Estimation of speaking style in speech corpora focusing on speech transcriptions. In *LREC*, pages 2747–2752.

Ingo Siegert, Ronald Böck, and Andreas Wendemuth. 2018. Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition. *Computer Speech & Language*, 51:1–23.

Deborah Tannen. 1980. The parameters of conversational style. In *18th Annual Meeting of the Association for Computational Linguistics*, pages 39–40.

Deborah Tannen. 1989. *That's Not What I Meant! How Conversational Style Makes or Breaks Relationships*. Ballantine.

Enrica Troiano, Aswathy Velutharambath, et al. 2021, submitted. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*.

Nigel G. Ward. 2019. *Prosodic Pattterns in English Conversation*. Cambridge University Press.

Nigel G. Ward. 2021a. Individual interaction styles: Evidence from a spoken chat corpus. In *SigDial*, pages 13–20.

Nigel G. Ward. 2021b. Interaction style variation: Companion website. Http://www.cs.utep.edu/nigel/istyles/.

Nigel G. Ward. 2021c. Interaction styles tools (2019-2022). https://github.com/nigelg ward/istyles.

Nigel G. Ward and Jonathan E. Avlia. 2022, submitted. A dimensional model of interaction style variation in spoken dialog. *Speech Communication*.

Richard Wright, Courtney Mansfield, and Laura Panfili. 2019. Voice quality types and uses in North American English. *Anglophonia*.

Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. A character expression model affecting spoken dialogue behaviors. In *Proceedings of the International Workshop on Spoken Dialog System Technology*.