Modeling Linguistic and Personality Adaptation for Natural Language Generation

Zhichao Hu[1], Jean E. Fox Tree[2] and Marilyn A. Walker[1]
Natural Language and Dialogue Systems Lab, Computer Science Department[1]
Spontaneous Communication Laboratory, Psychology Department[2]
University of California Santa Cruz, Santa Cruz, CA 95064, USA
{zhu, foxtree, mawalker}@ucsc.edu

Abstract

Previous work has shown that conversants adapt to many aspects of their partners' language. Other work has shown that while every person is unique, they often share general patterns of behavior. Theories of personality aim to explain these shared patterns, and studies have shown that many linguistic cues are correlated with personality traits. We propose an adaptation measure for adaptive natural language generation for dialogs that integrates the predictions of both personality theories and adaptation theories, that can be applied as a dialog unfolds, on a turn by turn basis. We show that our measure meets criteria for validity, and that adaptation varies according to corpora and task, speaker, and the set of features used to model it. We also produce fine-grained models according to the dialog segmentation or the speaker, and demonstrate the decaying trend of adaptation.

1 Introduction

Every person is unique, yet they often share general patterns of behavior. Theories of personality aim to explain these patterns in terms of personality traits, e.g. the Big Five traits of extraversion or agreeableness. Previous work has shown: (1) the language that people generate includes linguistic features that express these personality traits; (2) it is possible to train models to automatically recognize a person's personality from his language; and (3) it is possible to automatically train models for natural language generation that express personality traits (Pennebaker and King, 1999; Mairesse et al., 2007; Mairesse and Walker, 2011; Gill et al., 2012).

A distinct line of work has shown that people adapt to one another's conversational behaviors and that conversants reliably re-use or mimic many different aspects of their partner's verbal and non-verbal behaviors, including lexical and syntactical traits, accent, speech rate, pause length, etc. (Coupland et al., 1988; Willemyns et al., 1997; Brennan and Clark, 1996; Branigan et al., 2010; Coupland et al., 1988; Parent and Eskenazi, 2010; Reitter et al., 2006a; Chartrand and Bargh, 1999; Hu et al., 2014). Previous work primarily focuses on developing methods on measuring adaptation in dialog, and studies have shown that adaptation measures are correlated with task success (Reitter and Moore, 2007), and that social variables such as power affect adaptation (Danescu-Niculescu-Mizil et al., 2012).

Speaker (Utterance #): Utterance
F97: okay I'm on pacific avenue and plaza
D98: okay so you just take a right once your out of pacific lane you go wait no to late to your left.
F98: okay
D99: and I think. it's right ther- alright so I'm walking down pacific okay so it's right before the object it's right before the mission and pacific avenue intersection okay it's like umm almost brown and kinda like tan colored
F99: is it tan
D100: yeah it's like two different colors its like dark brown and orangey kinda like gold color its kinda like um
F100: okay is it kinda like a vase type of a thing
D101: yeah it has yeah like a vase

Figure 1: Dialog excerpt from the ArtWalk Corpus.

We posit that it is crucial to enable adaptation in computer agents in order to make them more human-like. However, we need models to control the amount of adaptation in natural language generation. A primary challenge is that dialogs exhibit many different types of linguistic features, any or all of which, in principle, could be adapted. Previous work has often focused on individual features when measuring adaptation, and referring expressions have often been the focus, but the conversants in the dialog in Figure 1 from the ArtWalk Corpus appear to be adapting to the discourse marker *okay* in D98 and F98, the hedge *kinda like* in F100, and to the adjectival phrase *like a vase* in D101.

Proceedings of the SIGDIAL 2018 Conference, pages 20–31,
Melbourne, Australia, 12-14 July 2018.

Therefore we propose a novel adaptation measure, Dialog Adaptation Score (DAS), which can model adaptation on any subset of linguistic features and can be applied on a turn by turn basis to any segment of dialog. Consider the example shown in Table 1, where the context (prime) is taken from an actual dialog. A response (target) with no adaptation makes the utterance stiff (DAS = 0), and too much adaptation (to all four discourse markers in prime, DAS = 1) makes the utterance unnatural. Our hypothesis is that we can learn models to approximate the appropriate amount of adaptation from the actual human response to the context (to discourse marker "okay", DAS = 0.25).

Conversants in dialogs express their own personality and adapt to their dialog partners simultaneously. Our measure of adaptation produces models for adaptive natural language generation (NLG) for dialogs that integrates the predictions of both personality theories and adaptation theories. NLGs need to operate as a dialog unfolds on a turn-by-turn basis, thus the requirements for a model of adaptation for NLG are different than simply measuring adaptation.

Context: *okay alright so yeah Im looking at 123 Locust right now*
Linguistic Features:
Discourse markers: okay, alright, so, yeah
Referring expressions: 123 Locust
Syntactic structures: `VP->VBP+VP`, `VP->VBG+PP+ADVB` ...

Adaptation Amount	Response	Adapted Features	DAS
None	*it should be somewhere*	None	0
Too much	*okay alright so yeah it should be somewhere*	okay, alright, so, yeah	1
Moderate	*okay I mean it should be somewhere*	okay	0.25

Table 1: Linguistic adaptation example: no adaptation, too much adaptation, and moderate adaptation (human response from ArtWalk Corpus).

We apply our method to multiple corpora to investigate how the dialog situation and speaker roles affect the level and type of adaptation to the other speaker. We show that:

- Different feature sets and conversational situations can have different adaptation models;
- Speakers usually adapt more when they have the initiative;
- The degree of adaptation may vary over the course of a dialog, and decreases as the adaptation window size increases.

2 Method and Overview

Our goal is an algorithm for adaptive natural language generation (NLG) that controls the system output at each step of the dialog. Our first aim therefore is a measure of dialog adaptation that can be applied on a turn by turn basis as a dialog unfolds. For this purpose, previous measures of dialog adaptation (Stenchikova and Stent, 2007; Danescu-Niculescu-Mizil et al., 2011) have two limitations: (1) their calculation require the complete dialog, and (2) they focus on single features and do not provide a model to control the interaction of multiple parameters in a single output, while our method measures adaptation with respect to any set of features. We further compare our method to existing measures in Section 6.

Measures of adaptation focus on prime-target pairs: (p, t), in which the prime contains linguistic features that the target may adapt to. While linguistic adaptation occur beyond the next turn, we simplify the calculation by using a window size of 1 for most experiments: for every utterance in the dialog (prime), we consider the next utterance by a different speaker as the target, if any. We show the decay of adaptation with increasing window size in a separate experiment. When generating (p, t) pairs, it is possible to consider only speaker A adapting to speaker B (target=A), only speaker B adapting to speaker A (target=B), or both at the same time (target=Both). In the following definition, $\mathrm{FC}_i(p)$ is the count of features in prime p of the i-th (p, t) pair, n is the total number of prime-target pairs in which $\mathrm{FC}_i(p) \neq 0$, similarly, $\mathrm{FC}_i(p \wedge t)$ is the count of features in both prime p and target t. We define Dialog Adaptation Score (DAS) as:

$$\mathrm{DAS} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{FC}_i(p \wedge t)}{\mathrm{FC}_i(p)}$$

Within a feature set, DAS reflects the average probability that features in prime are adapted in target across all prime-target pairs in a dialog. Thus our Dialog Adaptation Score (DAS) models adaptation with respect to feature sets, providing a whole-dialog adaptation model or a turn-by-turn adaptation model. The strength of DAS is the ability to model different classes of features related to individual differences such as personalities or social variables of interest such as status.

DAS scores measured using various feature sets can be used as a vector model to control adaptation in Natural Language Generation (NLG). Although

we leave the application of DAS to NLG to future work, here we describe how we expect to use it. We consider the use of DAS with three NLG architectures: Overgeneration and Rank, Statistical Parameterized NLG, and Neural NLG.

Overgenerate and Rank. In this approach, different modules propose a possibly large set of next utterances in parallel, which are then fed to a (trained) ranker that outputs the top-ranked utterance. Previous work on adaptation/alignment in NLG has made use of this architecture (Brockmann, 2009; Buschmeier et al., 2010). We can rank the generated responses based on the distances between their DAS vectors and learned DAS adaptation model. The response with the smallest distance is the response with the best amount of adaptation. We can also emphasize specific feature sets by giving weights to different dimensions of the vector and calculating weighted distance. For instance, in order to adapt more to personality and avoid too much lexical mimicry, one could prioritize related LIWC features, and adapt by using words from the same LIWC categories.

Statistical Parameterized NLG. Some NLG engines provide a list of parameters that can be controlled at generation time (Paiva and Evans, 2004; Lin and Walker, 2017). DAS scores can be used as generation decision probabilities. A DAS score of 0.48 for the LIWC feature set indicates that the probability of adapting to LIWC features in discourse context (prime) is 0.48. By mapping DAS scores to generation parameters, the generator could be directly controlled to exhibit the correct amount of adaptation for any feature set.

Neural NLG. Recent work in Neural NLG (NNLG) explores controlling stylistic variation in outputs using a vector to encode style parameters, possibly in combination with the use of a context vector to represent the dialog context (Ficler and Goldberg, 2017; Oraby et al., 2018). The vector based probabilities that are represented in the DAS adaptation model could be encoded into the context vector in NNLG. No other known adaptation measures could be used in this way.

We hypothesize that different conversational contexts may lead to more or less adaptive behavior, so we apply DAS on four human-human dialog corpora: two task-oriented dialog corpora that were designed to elicit adaptation (ArtWalk and Walking Around), one topic-centric spontaneous dialog corpus (Switchboard), and the MapTask Corpus used in much previous work. We obtain linguistic features using fully automatic annotation tools, described in Section 4. We learn models of adaptation from these dialogs on various feature sets. We first validate the DAS measure by showing that DAS distinguishes original dialogs from dialogs where the orders of the turns have been randomized. We then show how DAS varies as a function of the feature sets used and the dialog corpora. We also show how DAS can be used for fine-grained adaptation by applying DAS to individual dialog segments, and individual speakers, and illustrating the differences in adaptation as a function of these variables. Finally, we show how DAS scores decrease as the adaptation window size increases.

3 Corpora

We develop models of adaptation using DAS on the following four corpora.

ArtWalk Corpus (AWC).[1] Figure 1 provides a sample of the Artwalk Corpus (Liu et al., 2016), a collection of mobile-to-Skype conversations between friend and stranger dyads performing a real world-situated task that was designed to elicit adaptation behaviors. Every dialog involves a stationary director on campus, and a follower downtown. The director provided directions to help the follower find 10 public art pieces such as sculptures, mosaics, or murals in downtown Santa Cruz. The director had access to Google Earth views of the follower's route and a map with locations and pictures of art pieces. The corpus consists of transcripts of 24 friend and 24 stranger dyads (48 dialogs). In total, it contains approximately 185,000 words and 23,000 turns, from conversations that ranged from 24 to 55 minutes, or 197 to 691 turns. It includes referent negotiation, direction-giving, and small talk (non-task talk).[2]

Walking Around Corpus (WAC).[3] The Walking Around Corpus (Brennan et al., 2013) consists of spontaneous spoken dialogs produced by 36 pairs of people, collected in order to elicit adaptation behaviors, as illustrated by Figure 2. In each dialog, a director navigates a follower using a mobile phone to 18 destinations on a medium-sized campus. Directors have access to a digital map marked with

[1] https://nlds.soe.ucsc.edu/artwalk

[2] For AWC and WAC, we remove annotations such as speech overlap, noises (laugh, cough) and indicators for short pauses, leaving only clean text. If more than one consecutive dialog turn has the same speaker, we merge them into one dialog turn.

[3] https://catalog.ldc.upenn.edu/ldc2015s08

Speaker (Utterance #): Utterance
D137: and. you know on the uh other side of the math building like theres the uh, theres this weird, little concrete, structure that is sticking up out of the bricks, dont make any sense. F138: uh. D139: yeah youll see it when you get over there. F140: okay. D141: so just keep going and then uh. when you get around the building make a left. and you should be. F142: when I get around the Physics building make a left? D143: yeah yeah when you get around to the end here.

Figure 2: Dialog excerpt from the Walking Around Corpus.

target destinations, labels (e.g. "Ship sculpture"), photos and followers' real time location. Followers carry a cell phone with GPS, and a camera in order to take pictures of the destinations they visit. Each dialog ranges from 175 to 885 turns. The major differences between AWC and WAC are (1) in order to elicit novel referring expressions and possible linguistic adaptation, destinations in AWC do not have provided labels; (2) AWC happens in a more open world setting (downtown) compared to WAC (university campus).

Map Task Corpus (MPT).[4] The Map Task Corpus (Anderson et al., 1991) is a set of 128 cooperative task-oriented dialogs involving two participants. Each dialog ranges from 32 to 438 turns. A director and a follower sit opposite one another. Each has a paper map which the other cannot see (the maps are not identical). The director has a route marked on their map; the follower has no route. The participants' goal is to reproduce the director's route on the follower's map. All maps consist of line drawing landmarks labelled with their names, such as "parked van", "east lake", or "white mountain". Figure 3 shows an excerpt from the Map Task Corpus.

Switchboard Corpus (SWBD).[5] Switchboard (Godfrey et al., 1992) is a collection of two-speaker telephone conversations from all areas of the United States. An automatic operator handled the calls (giving recorded prompts, selecting and dialing another speaker, introducing discussion topics and recording the dialog). 70 topics were provided, for example: pets, child care, music, and buying a car. Each topic has a corresponding prompt message played to the first speaker, e.g. "find out what kind of pets the other caller has." A subset of 200K utterances of Switchboard have also been tagged with dialog act tags (Jurafsky et al., 1997). Each dialog contains 14 to 373 turns. Figure 1 provides an example of dialog act tags, such as `b` - Acknowledge (Backchannel), `sv` - Statement-opinion, `sd` - Statement-non-opinion, and `%` - Uninterpretable. We focus on this subset of the corpus.

Dialogs in SWBD have a different style from the three task-oriented, direction-giving corpora. Figure 4 illustrates how the SWBD dialogs are often lopsided: from utterance 14 to 18, speaker B states his opinion with verbose dialog turns, whereas speaker A only acknowledges and backchannels; from utterance 19 to 22, speaker A acts as the main speaker, whereas speaker B backchannels. Some theories of discourse define dialog turns as extending over backchannels, and we posit that this

Speaker (Utterance #): Utterance
D7: and below the graveyard below the graveyard but above the carved wooden pole. F8: oh hang on i don't have a graveyard. D9: okay. so you don't have a graveyard. do you have a fast flowing river. F10: fast running creek. D11: ehm mm don't know yeah it could be could be. F12: is that to the right that'll be to my right to my right. D13: to your. right uh-huh. F14: right. so i continue and go below the fast running creek. D15: no. go just until you go go below the diamond mine until just before the fast fast flowing river.

Figure 3: Dialog excerpt from the Map Task Corpus.

Speaker (Utterance #): [Tag] Utterance
B14: [b] Yeah. [sv] Well that's pretty good if you can do that. [sd] I know. [sd] I have a daughter who's ten [sd] and we haven't really put much away for her college up to this point [sd] but, uh, we're to the point now where our financial income is enough that we can consider putting some away A15: [b] Uh-huh. B16: [sd] for college [sd] so we are going to be starting a regular payroll deduction A17: [%] Um. B18: [sd] in the fall [sd] and then the money that I will be making this summer we'll be putting away for the college fund. A19: [ba] Um. Sounds good. [%] Yeah [sd] I guess we're, we're just at the point, uh [sd] my wife worked until we had a family [sd] and then, you know, now we're just going on the one income [sv] so it's B20: [b] Uh-huh. A21: [sv] a lot more interesting trying to, uh [sv] find some extra payroll deductions is probably the only way we will be able to, uh, do it. [sd] You know, kind of enforce the savings. B22: [b] Uh-huh.

Figure 4: Dialog excerpt from the Switchboard Dialog Act Corpus.

[4] http://groups.inf.ed.ac.uk/maptask/

[5] https://catalog.ldc.upenn.edu/ldc97s62

would allow us to measure adaptation more faithfully, so we utilize the SWBD dialog act tags to filter turns that only contain backchannels, keeping only dialog turns with tags `sd` (Statement-non-opinion), `sv` (Statement-opinion), and `bf` (Summarize/reformulate).[6] We then merge consecutive dialog turns from the same speaker.

4 Experimental Setup

We consider the following feature sets: unigram, bigram, referring expressions, hedges/discourse markers, and Linguistic Inquiry and Word Count (LIWC) features. Previous computational work on measuring linguistic adaptation in textual corpora have largely focused on lexical and syntactical features, which are included as baselines. Referring expressions and discourse markers are key features that are commonly studied for adaptation behaviors in task-oriented dialogs, which are often hand annotated. Here we automatically extract these features by rules. To model adaptation on the personality level, we draw features that correlate significantly with personality ratings from LIWC features. We hypothesize that our feature sets will demonstrate different adaptation models.

We lemmatize, POS tag and derive constituency structures using Stanford CoreNLP (Manning et al., 2014). We then extract the following linguistic features from annotations and raw text. The following example features are based on D137 in Figure 2.

Unigram Lemma/POS. We use lemma combined with POS tags to distinguish word senses. E.g., `lemmapos_building/NN` and `lemmapos_brick/NNS` in D137.

Bigram Lemma. E.g., `bigram_the-brick` and `bigram_side-of` in D137.

Syntactic Structure. Following Reitter et al. (2006b), we take all the subtrees from a constituency parse tree (excluding the leaf nodes that contain words) as features. E.g., `syntax_VP->VBP+PP` and `syntax_ADJP->DT+JJ` in D137. The difference is that we use Stanford Parser rather than hand annotations.

Referring Expression. Referring expressions are usually noun phrases. We start by taking all constituency subtrees with root `NP`, then map the subtrees to their actual phrases in the text and remove all articles from the phrase, e.g., `referexp_little-concrete` and `referexp_math-building` in D137.

Hedge/Discourse Marker. Hedges are mitigating words used to lessen the impact of an utterance, such as "actually" and "somewhat". Discourse markers are words or phrases that manage the flow and structure of discourse, such as "you know" and "I mean". We construct a dictionary of hedges and discourse markers, and use string matching to extract features, e.g., `hedge_you-know` and `hedge_like` in D137.

LIWC. Linguistic Inquiry and Word Count (Pennebaker et al., 2001) is a text analysis program that counts words in over 80 linguistic (e.g., pronouns, conjunctions), psychological (e.g., anger, positive emotion), and topical (e.g., leisure, money) categories. E.g., `liwc_second-person` and `liwc_informal` in D137. Because DAS features are binary, features such as Word Count and Number of New Lines are excluded.

Personality LIWC. Previous work reports for each LIWC feature whether it is significantly correlated with each Big Five trait (Mairesse et al., 2007) on conversational data (Mehl et al., 2006). For each trait, we create feature sets consisting of such features. See Table 2.

Personality	#	Example Features
Extraversion	15	Positive Emotion, Swear Words
Emotional Stability	14	Anger, Articles
Agreeable	16	Assent, Insight
Conscientious	17	Fillers, Nonfluencies
Open to Experience	12	Discrepancy, Tentative

Table 2: Number of LIWC features for each personality trait and example features.

5 Experiments on Modeling Adaptation

In this section, we apply our DAS measure on the corpora introduced in Section 3.

5.1 Validity Test: Original vs. Randomized Dialogs

We first establish that our novel DAS measure is valid by testing whether it can distinguish dialogs in their original order vs. dialogs with randomly scrambled turns (the order of dialog turns are randomized within speakers), inspired by similar approaches in previous work (Gandhe and Traum, 2008; Ward and Litman, 2007; Barzilay and Lapata, 2005). We calculate DAS scores for original dialogs and randomized dialogs using target=Both

[6] The filtering process removes 48.1% original dialog turns, but only 12.6% of the words. Filtered dialogs have 3 to 85 dialog turns each.

	#	Feature Sets	Original	Random
AWC	48	Unigram + Bigram	**0.10**	**0.07**
		All but LIWC	**0.13**	**0.10**
		LIWC	**0.48**	**0.46**
WAC	36	Unigram + Bigram	**0.22**	**0.19**
		All but LIWC	**0.18**	**0.16**
		LIWC	**0.55**	**0.54**
MPT	128	Unigram + Bigram	**0.27**	**0.24**
		All but LIWC	**0.20**	**0.18**
		LIWC	0.54	0.54
SWBD	1126	Unigram + Bigram	**0.18**	**0.17**
		All but LIWC	**0.20**	**0.19**
		LIWC	**0.67**	**0.66**

Table 3: Number of dialogs in four corpora, and average DAS scores of different feature sets for original and randomized dialogs. Bold numbers indicate statistically significant differences ($p < 0.0001$) between DAS scores for original and randomized dialogs in paired t-tests .

(Sec. 2) to obtain overall adaptation scores for both speakers.

We first test on lexical features (unigram and bigram) as in previous work. Then we add additional linguistic features (syntactic structure, referring expression, and discourse marker). These five features (see Section 4) are referred to as "all but LIWC". Finally, we test DAS validity using the higher level LIWC features.

We perform paired t-tests on DAS scores for original dialogs and DAS scores for randomized dialogs, pairing every original dialog with its randomized dialog. Table 3 shows the number of dialogs in each corpus, the average DAS scores of all dialogs within the corpus and p-values of corresponding t-tests. Although the differences between the average scores are relatively small, the differences in almost all paired t-tests are extremely statistically significant (cells in bold, $p < 0.0001$). The paired t-test on MPT using LIWC features shows a significant difference between the two test groups ($p < 0.05$). The original dialog corpora achieve higher average DAS scores than the randomized corpora for all 12 original-random pairs. The results show that DAS measure is sensitive to dialog turn order, as it should be if it is measuring dialog coherence and adaptation.

5.2 Adaptation across corpora and across features

This experiment aims to broadly examine the differences in adaptation across different corpora and feature sets. We first compute DAS on the whole dialog level for each feature set from Section 4, and then calculate the average across the corpus. We use target=Both (Sec 2) to obtain an overall measure of adaptation and leave calculating fine-grained DAS measures to Section 5.3. Table 4 provides results. We will refer to features in row 1 to 6 as "linguistic features" and row 7 to 11 as "personality features".

Row	Feature Sets	AWC	WAC	MPT	SWBD
1	Lemma/POS	0.14	0.15	0.29	0.28
2	Bigram	0.04	0.04	0.01	0.07
3	Syntax	0.17	0.14	0.11	0.28
4	ReferExp	0.03	0.03	0.01	0.01
5	Hedge	0.17	0.19	0.18	0.25
6	LIWC	0.48	0.55	0.53	0.71
7	Extra	0.40	0.46	0.30	0.58
8	Emot	0.48	0.50	0.38	0.72
9	Agree	0.47	0.51	0.44	0.71
10	Consc	0.38	0.44	0.20	0.55
11	Open	0.44	0.44	0.31	0.73

Table 4: Average DAS scores for each feature set.

Comparing columns, we first examine the DAS scores across different corpora. All p-values reported below are from paired t-tests. The two most similar corpora, the AWC and WAC, show no significant difference on linguistic features ($p = 0.43$). At the same time, the AWC and WAC do differ from the other two corpora. This demonstrates that the DAS reflects real similarities and differences across corpora. MPT shows lower DAS scores on all linguistic features except for lemma (word repetition), where it achieves the highest DAS score. With respect to personality features, WAC has significantly higher DAS scores than AWC ($p < 0.05$), possibly because of the different experiment settings: college student participants are more comfortable around their own campus than in downtown. MPT shows significantly lower DAS scores on personality features than AWC and WAC ($p < 0.05$). This may be because the MPT setting is the most constrained of the four corpora: being fixed in topic and location means dialogs are less likely to be influenced by environmental factors or to contain social chit chat. SWBD has the highest DAS scores in all feature sets except for referring expression. The higher DAS in non-referring features could be because the social chit chat allows more adaptation to occur. In addition, the dialogs we measure in SWBD are backchannel-filtered. The lower referring expression (respective to other SWBD scores) could be because SWBD does not require the referring expressions necessary

for the other three task-related corpora. We posit that the DAS adaptation models we present can be used in existing NLG architectures, described in Sec. 2. The AWC column in Table 4 shows adaptation model in the form of a DAS vector obtained from the ArtWalk Corpus.

Comparing rows, we then examine DAS scores among different features sets. LIWC has the highest DAS score among linguistic features, ranging from 0.48 to 0.71. While other linguistic features are largely content-specific, LIWC consists of higher level features that cover broader categories, thus its high DAS scores are expected. The DAS scores for the lemma feature range from 0.14 to 0.29, followed by Syntactic Structure (0.11 to 0.28), Hedge (0.17 to 0.25) and Bigram (0.01 to 0.07). Referring Expression has the lowest DAS score (0.01 to 0.03), possibly because our automatic extraction of referring expressions creates numerous subsets of one referring expression. Among personality features, Emotion Stability, Agreeableness, and Openness to Experience traits are adapted more than Extraversion and Conscientiousness. We leave to future work the question of why these traits have higher DAS scores.

5.3 Adaptation by Dialog Segment and Speaker

Our primary goal is to model adaptation at a fine-grained level in order to provide fine-grained control of an NLG engine. To that end, we report results for adaptation models on a per dialog-segment and per-speaker basis.

Reliable discourse segmentation is notoriously difficult (Passonneau and Litman, 1996), thus we heuristically divide each task-oriented dialog into segments based on number of destinations on the map: this effectively divides the dialog into subtasks. Since each dialog in SWBD only has one topic, we divide SWBD into 5 segments.[7] We compute DAS for each segment, and take an average across all dialogs in the corpus for each segment.

We compare all LIWC features vs. extraversion LIWC features because they provide high DAS scores across corpora. We also aim to explore the dynamics between two conversants on the extraversion scale. Figure 5 in Appendix illustrates how DAS varies as a function of speaker and dialog segment. In AWC, scores for all LIWC features slightly decrease as dialogs progress (Fig. 5(a)), while extraversion features show a distinct increasing trend with correlation coefficients ranging from 0.7 to 0.86 (Fig. 5(b)), despite being a subset of all LIWC features.[8] Average DAS displays the same decreasing trend in all and extraversion LIWC features for SWBD (Fig. 5(g) and 5(h)). We speculate that this might be due to the setup of SWBD: as the dialogs progress, conversants have less to discuss about the topic and are less interested. We also calculate per segment adaptation in WAC and MPT, but their DAS scores do not show overall trends across the length of the dialog (Fig. 5(c) to 5(f)).

We also explore whether speaker role and initiative affects adaptation. We use target=Both, target=D, and target=F to calculate DAS for each target.[9] We hypothesize that directors and followers adapt differently in task-oriented dialogs. In all task-oriented corpora (AWC, WAC, and MPT), we observe generally higher DAS scores with target=D, indicating that in order to drive the dialogs, directors adapt more to followers. In SWBD, the speaker initiating the call (who brings up the discussion topic and may therefore drive the conversation) generally exhibits more adaptation.

5.4 Adaptation on Different Window Sizes

This experiment aims to examine the trend of DAS scores as the window size increases. We begin with a window size of 1 and gradually increase it to 5. For a window size of n, the target utterance t is paired with the n-th utterance from a different speaker preceding t, if any. For example, in Figure 1, when window size is 3, target D100 is paired with prime F97; target D99 does not have any prime, thus no pair is formed.

Similar to Sec. 5.1, we compare DAS scores between dialogs in their original order vs. dialogs with randomly scrambled turns. We hypothesize that similar to the results of repetition decay measures (Reitter et al., 2006a; Ward and Litman, 2007; Pietsch et al., 2012), the DAS scores of original dialogs would decrease as the window size increases. We use target=both to obtain overall adaptation scores involving both speakers, and calculate DAS with all but the Personality LIWC feature sets introduced in Sec. 4. We first compute DAS on the whole dialog level for each window size, and then calculate the average DAS for each window size

[7] To ensure two way adaptation exists in every segment (both speaker A adapting to B, and B adapting to A), the minimum length (number of turns) of each segment is 3. Thus we only work with dialogs longer than 15 turns in SWBD.

[8] Using Simple Linear Regression in Weka 3.8.1.

[9] In task-oriented dialogs, D stands for Director, F for Follower. In SWBD, D stands for the speaker initiating the call.

across the corpus.

Results show that DAS scores for the original dialogs in all corpora decrease as window size increases, while DAS scores for the randomized dialogs stay relatively stable. Figure 6 in Appendix shows plots of average DAS scores on different window sizes for original and randomized dialogs. Plots of the AWC and WAC show similar trends. Experiments with larger window sizes show that the original and random scores meet at window size 6 - 7 (with different versions of randomized dialogs). In MapTask, the original and random scores meet at window size 3 - 4. In SWBD, original and random scores meet at window size 2.

6 Related Work

Recent measures of linguistic adaptation fall into three categories: probabilistic measures, repetition decay measures, and document similarity measures (Xu and Reitter, 2015). Probabilistic measures compute the probability of a single linguistic feature appearing in the target after its appearance in the prime. Some measures in this category focus more on comparing adaptation amongst features and do not handle turn by turn adaptation (Church, 2000; Stenchikova and Stent, 2007). Moreover, these measures produce scores for individual features, which need aggregation to reflect overall adaptivity (Danescu-Niculescu-Mizil et al., 2011, 2012). Document similarity measures calculate the similarity between prime and target by measuring the number of features that appear in both prime and target, normalized by the size of the two text sets (Wang et al., 2014). Both probabilistic measures and document similarity measures require the whole dialog to be complete before calculation.

Repetition decay measures observe the decay rate of repetition probability of linguistic features. Previous work has fit the probability of linguistic feature repetition decrease with the distance between prime and target in logarithmic decay models (Reitter et al., 2006a,b; Reitter, 2008), linear decay models (Ward and Litman, 2007), and exponential decay models (Pietsch et al., 2012).

Previous work on linguistic adaptation in natural language generation has also attempted to use adaptation models learned from human conversations. The alignment-capable microplanner SPUD *prime* (Buschmeier et al., 2009, 2010) uses the repetition decay model from Reitter (2008) as part of the activation functions for linguistic structures. However, the parameters are not learned from real data. Repetition decay models do well in statistical parameterized NLG, but is hard to apply to overgenerate and rank NLG. Isard et al. (2006) apply a pre-trained n-grams adaptation model to generate conversations. Hu et al. (2014) explore the effects of adaptation to various features by human evaluations, but their generator is not capable of deciding which features to adapt based on input context. Dušek and Jurčíček (2016) use a seq2seq model to generate responses adapting to previous context. They utilize an n-gram match ranker that promotes outputs with phrase overlap with context. Our learned adaptation models could serve as a ranker. In addition to n-grams, DAS could produce models with any combinations of feature sets, providing more versatile adaptation behavior.

7 Discussion and Future Work

To obtain models of linguistic adaptation, most measures could only measure an individual feature at a time, and need the whole dialog to calculate the measure (Church, 2000; Stenchikova and Stent, 2007; Danescu-Niculescu-Mizil et al., 2012; Pietsch et al., 2012; Reitter et al., 2006b; Ward and Litman, 2007). This paper proposes the Dialog Adaptation Score (DAS) measure, which can be applied to NLG because it can be calculated on any segment of a dialog, and for any feature set.

We first validate our measure by showing that the average DAS of original dialogs is significantly higher than randomized dialogs, indicating that it is sensitive to dialog priming as intended. We then use DAS to show that feature sets such as LIWC, Syntactic Structure, and Hedge/Discourse Marker are adapted more than Bigram and Referring Expressions. We also demonstrate how we can use DAS to develop fine-grained models of adaptation: e.g. DAS applied to model adaptation in extraversion displays a distinct trend compared to all LIWC features in the task-oriented dialog corpus AWC. Finally, we show that the degree of adaptation decreases as the window size increases. We leave to future work the implementation and evaluation of DAS adaptation models in natural language generation systems.

Acknowledgement

This research was supported by NSF CISE RI EAGER #IIS-1044693, NSF CISE CreativeIT #IIS-1002921, NSF CHS #IIS-1115742, Nuance Foundation Grant SC-14-74, and auxiliary REU supplements.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech* 34(4):351–366.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 331–338.

H.P. Branigan, M.J. Pickering, J. Pearson, and J.F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42(9):2355–2368.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(6):1482.

Susan E Brennan, Katharina S Schuhmann, and Karla M Batres. 2013. Entrainment on the move and in the lab: The walking around corpus. In *Proc. of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.

Carsten Brockmann. 2009. *Personality and Alignment Processes in Dialogue: Towards a Lexically-Based Unified Model*. Ph.D. thesis, University of Edinburgh, School of Informatics.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*. Association for Computational Linguistics, pages 82–89.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2010. Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. *Empirical methods in natural language generation* 5980.

Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology* 76(6):893.

Kenneth W Church. 2000. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p 2. In *Proc. of the 18th conference on Computational linguistics-Volume 1*. pages 180–186.

N. Coupland, J. Coupland, H. Giles, and K. Henwood. 1988. Accommodating the elderly: Invoking and extending a theory. *Language in Society* 17(1):1–41.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*. ACM, pages 745–754.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*. ACM, pages 699–708.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the SIGDIAL 2016 Conference*. Association for Computational Linguistics, pages 185–190.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633* .

Sudeep Gandhe and David Traum. 2008. An evaluation understudy for dialogue coherence models. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, pages 172–181.

Alastair J Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *Proc. of the Seventh International Natural Language Generation Conference*. pages 40–48.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, volume 1, pages 517–520.

Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. 2014. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, Springer, pages 151–164.

Amy Isard, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and alignment in generated dialogues. In *Proceedings of the Fourth International Natural Language Generation Conference*. Association for Computational Linguistics, pages 25–32.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report* pages 97–102.

Grace Lin and Marilyn Walker. 2017. Stylistic variation in television dialogue for natural language generation. In *EMNLP Workshop on Stylistic Variation*.

Kris Liu, Jean E Fox Tree, and Marilyn A Walker. 2016. Coordinating communication in the wild: The artwalk dialogue corpus of pedestrian navigation and mobile referential communication. In *LREC*.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)* 30:457–500.

Francois Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* .

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology* 90:862–877.

Shereen Oraby, Lena Reed, Shubhanghi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the SIGDIAL 2018 Conference: The 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics.

Daniel S. Paiva and Roger Evans. 2004. A framework for stylistically controlled generation. In Anja Belz, Roger Evans, and Paul Piwek, editors, *Natural Language Generation, Third Internatonal Conference, INLG 2004*. Springer, number 3123 in LNAI, pages 120–129.

Gabriel Parent and Maxine Eskenazi. 2010. Lexical entrainment of real users in the lets go spoken dialog system. In *Proceedings Interspeech*. pages 3018–3021.

Rebecca J. Passonneau and Diane Litman. 1996. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence and linguistic devices. In Donia Scott and Eduard Hovy, editors, *Computational and Conversational Discourse: Burning Issues - An Interdisciplinary Account*, Springer-Verlag, Heidelberg, Germany, pages 161–194.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ.

J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77:1296–1312.

Christian Pietsch, Armin Buch, Stefan Kopp, and Jan de Ruiter. 2012. Measuring syntactic priming in dialogue corpora. *Empirical Approaches to Linguistic Theory: Studies in Meaning and Structure* 111:29.

David Reitter. 2008. *Context effects in language production: models of syntactic priming in dialogue corpora*. Ph.D. thesis, University of Edinburgh. http://www.david-reitter.com/pub/reitter2008phd.pdf.

David Reitter, Frank Keller, and Johanna D Moore. 2006a. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, pages 121–124.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Annual Meeting-Association for Computational Linguistics*. volume 45, page 808.

David Reitter, Johanna D. Moore, and Frank Keller. 2006b. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Vancouver, Canada, pages 685–690. http://www.david-reitter.com/pub/reitter2006priming.pdf.

Svetlana Stenchikova and Amanda Stent. 2007. Measuring adaptation between dialogs. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*.

Yafei Wang, David Reitter, and John Yen. 2014. Linguistic adaptation in online conversation threads: analyzing alignment in online health communities. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics (at ACL)*. Baltimore, Maryland, USA, pages 55–62. http://www.david-reitter.com/pub/yafei2014cmcl.pdf.

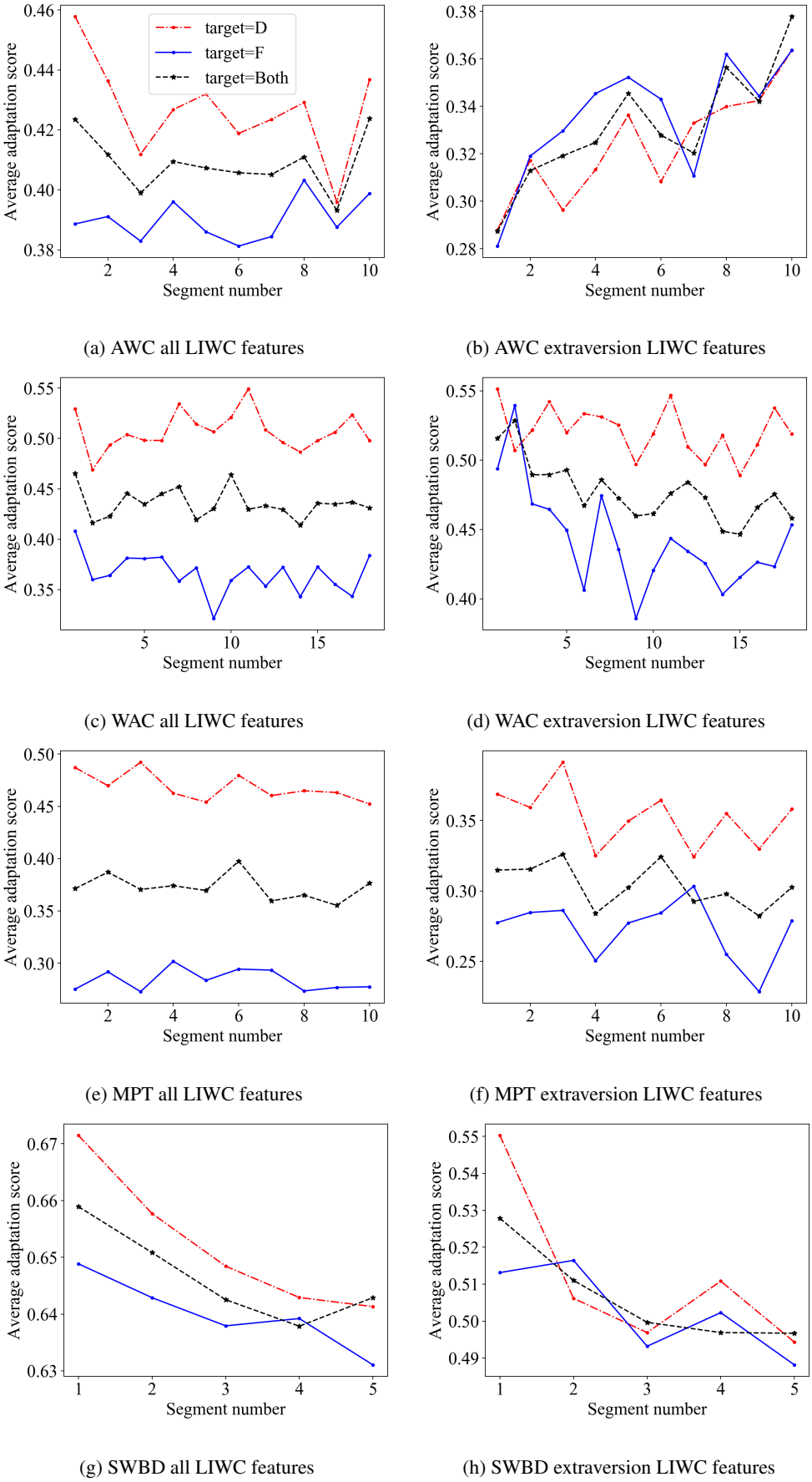
Arthur Ward and Diane Litman. 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proc. of the SLaTE Workshop on Speech and Language Technology in Education*.

Michael Willemyns, Cynthia Gallois, Victor J Callan, and Jeffery Pittam. 1997. Accent accommodation in the job interview impact of interviewer accent and gender. *Journal of Language and Social Psychology* 16(1):3–22.

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proc. Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Denver, CO, pages 58–67. http://www.david-reitter.com/pub/xu2015evaluation-alignment.pdf.

Appendix

Figure 5 and Figure 6.



(a) AWC all LIWC features

(b) AWC extraversion LIWC features

(c) WAC all LIWC features

(d) WAC extraversion LIWC features

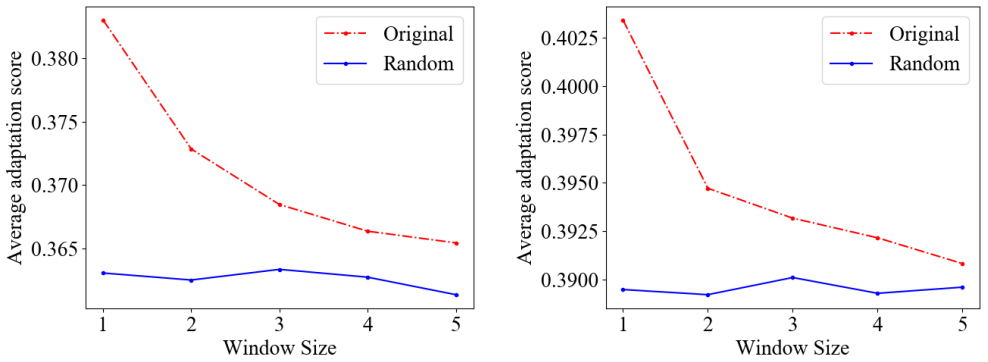
(e) MPT all LIWC features

(f) MPT extraversion LIWC features

(g) SWBD all LIWC features

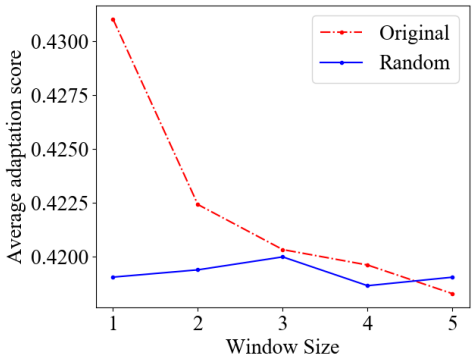
(h) SWBD extraversion LIWC features

Figure 5: Plots of average DAS as the dialogs progress, using all LIWC features vs. extraversion LIWC features.

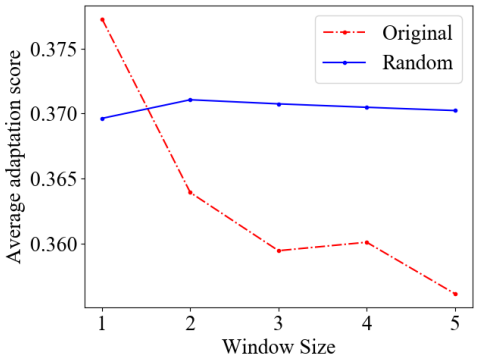


(a) ArtWalk Corpus

(b) Walking Around Corpus



(c) MapTask Corpus



(d) Filtered Switchboard Corpus

Figure 6: Plots of average DAS on different window sizes (1 to 5) for original dialogs vs. randomized dialogs, using all feature sets except Personality LIWC.