The contribution of dialogue act labels for convergence studies in natural conversations

Simone Fuscone[1,2], **Benoit Favre**[2] and **Laurent Prévot**[1,3]
[1] Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2] Aix Marseille Univ, CNRS, LIS, Marseille, France
[3] Institut Universitaire de France, Paris, France

Abstract

Speakers convergence in conversation had been studied because of its relevance for cognitive models of communication as well as for dialogue systems adaptation to the user. Convergence effects have been established on controlled datasets. Tracking interpersonal dynamics on generic corpora has provided positive but more contrasted outcomes. We propose here to enrich large conversational corpora with dialogue act information and to use acts as filters to create sub sets featuring homogeneous conversational activity. Those sub sets allow a more precise comparison between speakers' speech variables. Our experiences consist of comparing convergence on low level variables (Energy, Pitch, Speech Rate) measured on raw datasets, with human and automatically labelled datasets. We found that such filtering does help in observing convergence suggesting that future studies should consider such high level dialogue activity types and the related NLP tools as important aspects for analyzing conversational interpersonal dynamics.

1 Introduction

Individual speech characteristics of the participants engaged in a conversation depend on speech activities of both participants and on their interaction. This intuition has taken shape in numerous studies and even resulted in general models of communication as described in the *accommodation theory* (Giles et al., 1991) and in the *interactive alignment* model (Pickering and Garrod, 2004). The phenomenon has been studied under different angles such as speech or phonetic *convergence* (Street, 1984; Pardo, 2006), prosodic *entrainment* (Levitan and Hirschberg, 2011; Truong and Heylen, 2012).

Earlier work from (Edlund et al., 2009) found that participants to a conversation tend to be more similar (in terms of gaps and pauses duration) to their partner than chance would predict. This is in line with the more recent study on speech rate from (Cohen Priva et al., 2017). However, the absence of significant results in comparing the inter-speaker distance in the first and second halves of the conversation make them conclude that convergence "cannot be captured by sampling the distance between speakers at two different stages of the dialogue". The study in (Truong and Heylen, 2012) conducted a similar experiment (investigating intensity and pitch) on English MapTask (Anderson et al., 1991) with little results as well and conclude that "the measure of alignment remains a complicated matter due to its dynamic nature and to the social factors that can influence the amount of convergence". Although the literature has proposed many different ways of approaching these questions, they all rely on extracting features from a certain period or point in time and comparing them.

In this paper we would like to push further the investigation of interpersonal dynamics in real-life corpora. Considered from the angle of speech and linguistic variables, an essential aspect of conversational corpora is their huge variability. This variability is due to a large extent to the different conversational activities speakers can participate in. For instance, they can enter in a storytelling sequence in which one interlocutor become less active, enter into a heated debate, etc... We propose here to use the latest dialogue tagging techniques to create subsets obtained from dialogue acts filtering in which frequent dialogue acts are used as a proxy to characterize the conversational activity of a given turn, and therefore compare data points that are more homogeneous.

Our approach combines three ingredients that were not used together in the past and current literature to our best knowledge. First, we recognize

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020, Online, hosted from Massachusetts, USA.

the social factor and hypothesis that the lower level variables (such as energy) are less prone to social factor or strategic adaptation that other variables. This is justified by the fact that production and perception processes are deeply linked in conversation and that entrainment at those lower level seems to be automatic. Second, we consider that some amount of convergence should be observed within the time frame of a conversation even though we acknowledge that the "two halves" approach is extremely crude in this respect. Third, similarly to (Cohen Priva et al., 2017) our approach is based on a large conversational corpus with the intention of overcoming the noise and the small size of the effect by increasing the amount of data considered.

The paper starts with a review (Section 2) of related works describing previous studies that focus on convergence. Then in Section 3 we introduce the dataset that we used in our analysis and we describe the DA tagger. Section 4 shows the feature extraction of the variables we scrutinized in this work and the methodology we apply to measure convergence. The results are illustrated in Section 5. Finally, we discuss the results and possible future improvements and open aspects in Section 6.

2 Related work

Conversational Interpersonal Dynamics have been approached at different granularity levels and for a large range of variables. In terms of granularity, studies can be (i) **Inter-conversation** comparisons ; (ii) **Intra-conversation** (focusing on the dynamics within conversations). Inter-conversation comparison could be simple correlation studies between the speakers (Edlund et al., 2009) or, when the data allows, comparison between values of a speaker and its conversational partners vs. a speaker and all other non-partner corpus participants (Cohen Priva et al., 2017). Intra-conversation studies vary a lot in term of approaches ranging from coarse-grained "difference-in-difference" *convergence* (Edlund et al., 2009; Truong and Heylen, 2012; Cohen Priva and Sanker, 2018) approaches consisting in comparing differences between speakers in different intervals to fine-grained *synchrony* approaches using sliding window in order to compare local speaker similarities (Truong and Heylen, 2012).

While a large body of carefully controlled experiments on lab speech provided results on convergence, the results on real corpora (from the studies listed in the previous paragraph) provide a more complex pictures, the exhibition of some effect for some variables but with a relative fragility of the magnitude of these effects (Fuscone et al., 2018) and arise overall many comments on the methodological subtleties and mentions to room for improvements (See (Truong and Heylen, 2012; Cohen Priva and Sanker, 2018) for instance).

Finally in regards to the scrutinized features of speech, (Natale, 1975) has shown that modulating experimentally the intensity of voice of a shadow partner during the conversation led the participants to change their intensity converging to the partner's values. (Edlund et al., 2009) targeted interactional variables of pauses and inter-speakers gaps duration showing convergence between the speakers. Pitch was the focus of (Truong and Heylen, 2012; Bonin et al., 2013) while speech rate had been investigated for example in (Cohen Priva et al., 2017).

Our hypothesis, derived from both the empirical results above and the theoretical models (Pickering and Branigan, 1998; Giles and Powesland, 1997), is that automatic *entrainment* and strategic *adaptation* are blending in to produce *convergence* and *synchrony* phenomena. This blend is complex but intuitively, respectively low-level variables (such as intensity) and high-level variables (such as lexical or syntactic choices) may be more impacted by automatic entrainment and strategic adaptation. This could explain why firmer and more results seems to be obtained on low-level variables (Natale, 1975; Levitan, 2014).

To summarize conversational interpersonal dynamics, that is supposed to be an established phenomenon, can be surprisingly difficult to track in real conversations. The main issue is the heterogeneity of the speech activities both within a conversation (huge and rapid variation across time depending on what interaction is doing) and across conversations. We propose here to use dialogue acts to "organize" and filter conversational activities within large generic corpora. Moreover, to account for *adaptation* one must take precise care of speaker profiles. Our approach therefore focuses on relatively low level variable to avoid as much as possible the "adaptation" part of the interpersonal dynamics.

Our question is whether we can observe more reliably interpersonal dynamics in raw, manually

	All	St.	Opi.	Bc.
SWBD	180h	-	-	-
SWDA	41h	17h	7 h	1h
SW-Auto	119h	47h	19h	3h

Table 1: Results Summary : Hours of speech excluding noise and silence

DA-tagged (smaller) or automatically DA-tagged (larger) dataset even if we reduce the data size. An underlying question being whether the noise introduced by the DA-tagging uncertainty and the data size reduction is compensated by the gain in homogeneity between the material that is compared. Finally, a side question is whether some DA-specific subsets (e.g backchannel) are interesting sub datasets for studying specific interpersonal effects.

3 Dialogue Act Filtering and DataSets

In conversational speech, a large amount of variability that can be observed for a given speech variable across time is due to the different speech activities that are performed by the speakers. For instance, participants can remain rather passive, be very narrative or inquisitive, etc.; We attempt to deal with this crucial issue by using dialogue acts (DA) as a proxy for creating subset corresponding to different "dialogue activity". We only used the acts dominating the DA distribution: *Statement, Opinion, Backchannel*

3.1 Raw dataset

Switchboard (Godfrey et al., 1992) is a corpus of telephonic conversations between randomly assigned speakers[1] of American English discussing a preassigned topic. The corpus consists of 2430 conversations (of an average duration of 6 minutes) for a total of 260 hours, involving 543 speakers. The corpus has audio, time aligned transcripts and a segmentation into *utterances*.

3.2 Manually tagged datasets

642 Switchboard conversations have been segmented and annotated for dialogue acts (DA) (Calhoun et al., 2010). The DA-tagged set has been simplified to 42 tags but a few of them (Statement: 36%, Backchannel: 19%, Opinion: 13%) are dominating the distribution, illustrated in Table 2. See (Stolcke et al., 1998) for the details.

[1]Speakers therefore do not know each other.

DA type	**Example**
STA	*"And that was pretty heartrending for her"*
OPI	*"money seems to be too big of an issue."*
BAC	*"Uh-huh."*
AGR	*"you're right"*

Table 2: Examples for the DA types used.

	Prediction Score		
Class	Precision	Recall	F1
BAC	0.78	0.84	0.81
STA	0.78	0.73	0.75
OPI	0.47	0.61	0.53
OTH	0.71	0.68	0.69

Table 3: Prediction score of the Turn Tagger for the classes: Backchannel (BAC), Statement (STA), Opinion (OPI).

3.3 Automatically tagged dataset

Since Switchboard provides human annotation for a subset of the entire corpus, we propose to use a **turn tagger** on each conversation.

We considered to first try the categories that contain the majority of classes, said *Statement* (STA), *Backchannel* (BAC), *Opinion* (OPI) and *Other* (OTH) as shown in Figure 1. We used as train, development and test sets the NXT Switchboard corpus that contains annotated Dialog Acts for 642 conversations. As the dialog acts do not match the turn segmentation of the conversations, we label each turn of the corpus by assigning one of the majority class, among the DA tags used in the turn. The model we used is described in (Auguste et al., 2018) and inspired by the model of (Yang et al., 2016). It is a two level hierarchical Neural Network (with learning rate = 0.001, batch size = 32, max length of each turn = 80, embeddings words dimension = 200). In the first level each turn is treated singularly taking into account the words that form the turn while the second level is used to take into account the whole turn in the context of the conversation. Each level is a bidirectional Long Short Term Memory (LSTM). We used 80% of switchboard data as training set, 10% for development and 10% for the test set.

Score for the classes are reported in Table 3. In particular the class Opinion presents a low score in both precision and recall. As expected the class Other as well, that contains all other DA types, has a low score. In order to improve

	Prediction Score		
Class	Precision	Recall	F1
BAC+AGR	0.88	0.85	0.86
STA+OPI	0.84	0.92	0.87
OTH	0.62	0.49	0.55

Table 4: Prediction score of the Turn Tagger for the 3 classes.

the precision and recall of the tagger, we try to reduce the noise of the tagging task by adopting a grouping of similar categories, corresponding to *Statement+Opinion* (STA+OPI), *Backchannel+Agreement* (BAC+AGR) and *Other* (OTH) which includes all the other dialog acts. This grouping was obtain by first considering only the DA dominating the distribution. Then we manual inspect many examples of each dialogue act and figure that although functionally different *statements* and *opinions* on the one hand; *backchannel* and *Agreement* on the other hand corresponded to very similarly conversational activities. More precisely, the former have clear *main speaker* feeling with a lot of semantic content while the later have a much more *listener* nature with various kinds of feedback related lexical items (see Table 2). The resulting distribution is 52% STA+OPI, 25% BAC+AGR and 23% of OTH. The Accuracy of the DA tagger is 81% on the test set, the details for each category is reported in table 4. The accuracy of the class OTH, as expected, is lower compared to the other 2 classes, taking into account that it is formed by heterogeneous DA acts.

4 Analysis

4.1 Feature processing

Energy (E) and **Pitch (P)** are computed from the audio files with the help of *openSMILE* (feature extraction and audio analysis tool) by (Eyben and Schuller, 2015) while **Speech Rate (R)** is computed using time aligned transcripts.

Energy (E): One of the issue of telephonic conversation is the distance mouth-microphone that affects measured values of voice intensity. This adds noise to the distribution of values across speakers, even for the same speaker across different conversations. So to reduce this effect we introduce a normalization factor consisting in dividing each speaker E values by the average E produced by that speaker in the entire conversation. In addition, to reduce the environmental noise, we computed the average E using the temporal windows where the probability of voicing is above 0.65. Then we computed for each conversational unit (turn or Dialog act as provided by Switchboard transcripts) the average E.

Pitch (P): We computed the average in each conversational unit for each speaker.

Speech Rate (R): We used the approach proposed by Cohen-Priva (Cohen Priva et al., 2017) that defines R for an utterance as the ratio between the actual duration of the utterance and its expected duration (computed by estimating every word duration into the whole corpus, for all speakers). Values above / below 1 correspond respectively to fast / slow speech compare to the average of the corpus. In order to make the measure R more reliable we consider only utterances having more than 5 tokens.

4.2 Convergence

We divide each conversation into two halves and compare the distance between the average values of the target variables coming from each speaker. (Truong and Heylen, 2012; Edlund et al., 2009).

We computed the difference between the mean value of target variables for the two speakers in both halves. This provided us two values (first and second interval) for each variable and each conversation:

$$\Delta\overline{V}_i = \mid \overline{V}_{Ai} - \overline{V}_{Bi} \mid \quad (1)$$

, where $i = 1, 2$ refers respectively to the first and second interval, A and B indicate the speakers who take part in the conversation while V could be E (Energy), $F0$ (Pitch) and R (Speech rate). Our aim is to test the hypothesis that convergence, on such rather low-level variables, occurs during the interaction. We therefore computed the difference between both intervals, resulting in a distribution of these values in both intervals for the whole corpus. We then fitted a linear mixed regression model to this distribution to test if there is a significant difference across the intervals. Moreover, the sign of the estimate of the model provides us the direction of the evolution. We use the `lme4` library in `R` (Bates et al., 2014) to fit the models and provide t-values. The `lmerTest` package (Kuznetsova et al., 2014), which encapsulates `lme4`, was used to estimate degrees of freedom. (*Satterthwaite approximation*) and calculate p-values. In the model, the $\Delta\overline{V}_i$ is the predicted value, the A and B identities as well as the topic of the conversation are set

	All	St.	Opi.	Bc.
SWBD	**E**R	X	X	X
SW-DA-Man.	ER	ER	-	E
SW-DA-Auto.	X	-	-	E

Table 5: Difference-In-Difference Results Summary: E: Energy; P: Pitch ; R: Speech Rate; -: any significance; X: no performed experiments. normal font : p-value$\leq$0.05 ; **bold** : p-value $\leq$ 0.01

as random intercepts. The model, in R notation, is $\Delta\overline{V}_i \sim t_i + (1 \mid topic) + (1 \mid speaker_A) + (1 \mid speaker_B)$.

We computed the mean target variables in the two intervals following the same method as for the whole dataset. Excluding conversations with undefined values (miss one or more average interval values), the number of conversations is 593 for *Statement*, 581 for *Backchannel* and 381 for *Opinion*.

5 Results

We report the results in the case of the whole dataset without DA (SWBD) and the DA manually tagged (SW-DA-Man.), the results for the Dialog acts grouping that refers to SW-DA-Man. and the dialog acts automatically tagged (SW-DA-Auto). The results are summarized in Tables 5 for Inter-Speaker correlations and 6 for convergence.

	All	St. + Opi.	Bc. +Agr.
SWBD	**E**R	X	X
SWBD-Auto	X	ER	E

Table 6: Difference-In-Difference Results Summary: E: Energy; P: Pitch ; R: Speech Rate; -: any significance; X: no performed experiments. normal font : p-value$\leq$0.05 ; **bold** : p-value $\leq$ 0.01

5.1 No Dialog acts filtering (SWBD and SW-DA-Man.)

The use of the whole dataset without considering DA filtering is shown in Table 7. Convergence arises for Energy and Speech Rate. When using the DA-tagged part of Switchboard (642 conversations, corresponding to 41 hours when excluding silence and noise) as reported in table 8, there is a weaker effect both on E and R while F0 doesn't show any significant effect. The less strong significance compared to the SWBD case is expected, considering the huge data size reduction. For R instead, the effect seems to be stronger in this case despite the reduction of conversations.

SWBD	*Entire Corpus (180 hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.063**	**0.012**	$\mathbf{7 \times 10^{-7}}$
P-Mean	-0.044	0.021	0.490
R-Mean	**-0.049**	**0.024**	**0.046**

Table 7: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for backchannel.

SW-DA-Man.	*Whole DA-tagged subset (41 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.054**	**0.021**	**0.026**
P-Mean	-0.057	0.040	0.158
R-Mean	**-0.106**	**0.047**	**0.026**

Table 8: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for backchannel.

5.2 DA filtering (SW-DA-Man.)

The results for the categories statement, opinion and backchannel are reported in Table 9.

The *backchannel* subset exhibits convergence for E despite the huge data size reduction. However, no significant effects in F0 and R were found in category most likely because *backchannel* are made of very short utterances, including frequently one word and for which estimated duration is a problematic question.

Opinion DA-tag filtering seems to reduce too much both the data size as well as the samples number participating to the mean calculation for each halves and provides more noisy data in which significance cannot be attained.

The *statement* subset shows convergence for energy and speech rate. Considering just statement seems to clean the dataset by feedback-related differences as well as strong disfluencies (type *abandoned* in Switchboard). This helps observing the effect for speech rate. Contrarily, the wide variety of *statements* in terms of utterance duration could be an issue for pitch since contours and physiological-related decreasing slope could result in a lot of noise for this variable. Overall the comparison of the results of the non-filtered datasets and of the

SW-DA-Man.	*Backchannel (1 Hour)*		
Feature	Estimate	std	p-values
E-Mean	**-0.082**	**0.041**	**0.045**
P-Mean	0.043	0.022	0.491
SW-DA-Man.	*Statement (17 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.071**	**0.023**	**0.032**
P-Mean	-0.025	0.038	0.653
R-Mean	**-0.123**	**0.049**	**0.012**
SW-DA-Man.	*Opinion (7 Hours)*		
Feature	Estimate	std	p-values
E-Mean	-0.061	0.033	0.627
P-Mean	-0.032	0.053	0.552
R-Mean	-0.096	0.061	0.115

Table 9: Results for **SW-DA manually tagged**: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for back-channels.

SW-DA-Auto	*Backchannel (3 Hours)*		
CLASS	Estimate	std	p-values
E-mean	**-0.079**	**0.031**	**0.009**
P-mean	0.003	0.002	0.815
SW-DA-Auto	*Statement (47 Hours)*		
CLASS	Estimate	std	p-values
E-mean	-0.040	0.025	0.2396
P-mean	0.009	0.004	0.611
R-mean	-0.004	0.004	0.332
SW-DA-Auto	*Opinion (19 Hours)*		
CLASS	Estimate	std	p-values
E-mean	0.011	0.048	0.811
P-mean	-0.014	0.032	0.663
R-mean	-0.008	0.009	0.369

Table 10: Results for **SW-DA automatically tagged**: Parameters of our linear model for energy, pitch and speech rate for the corpus automatically tagged considered statement, opinion and backchannel as DA classes. Speech rate was not considered for backchannel

filtered subsets suggests interesting patterns. For whole dataset, the data on which we compute the mean is much more dense and therefore the calculation robust, this seems to be enough to establish a trend for E and R. However, F0 doesn't show any significant effect for all the dataset (entire, subset and DA-acts of the Switchboard corpus). This probably can be explained considering that pitch is a more complex variable in the natural conversations framework and an average approach can't get the more strategic behavior of pitch ((Reichel et al., 2018)).

Backchannel is a strong filter (that controls for lexical content, duration and even the utterance function which is precisely delineated – the most frequent items are "Yeah" $= 35\%$, "Uh-huh" $= 18\%$, "Um-hum" $= 16\%$, "Right" $= 9\%$). However estimates for speech rate for *backchannel* is extremely problematic because of its linguistic form. In the case of *opinion* the strong reduction of data size is likely the cause to not find any effect. Finally, about *statements* it seems that filtering out very different utterances helps in bringing a more coherent dataset and therefore observing the convergence effect on E and R.

5.3 DA automatically tagged (SW-DA-Auto)

Here we report the results we obtained using the DA, grouped as statement, opinion and backchannel in Table 10 and the grouping of statement + opinion, backchannel + agreement in Table 11.

Comparing the results of the dataset automatically tagged and the subset with human annotation 5.2 we note that the only variable that shows convergence is energy in the case of *backchannel*. Even the number of conversations increases compared to the manual annotated dataset, the tagging stage could introduce noise derived by the turn that are non correctly labeled. As consequence this could have been affected the measure of distances between speakers.

Using indeed the classes derived by grouping similar classes (statement + opinion and backchannel + agreement) results in an emerging conver-

SWDA-Auto	*Backchannel + Agreement*		
CLASS	Estimate	std	p-values
E-mean	**-0.079**	**0.028**	**0.006**
P-mean	0.053	0.028	0.192
SWDA-Auto	*Statement + Opinion*		
CLASS	Estimate	std	p-values
E-mean	**-0.055**	**0.011**	$\mathbf{4 \cdot 10^{-6}}$
P-mean	-0.035	0.038	0.353
R-mean	**-0.075**	**0.021**	**0.008**

Table 11: Parameters of our linear model for energy, pitch and speech rate for the corpus automatically tagged considered statement + opinion and backchannel + agreement as DA classes. Speech rate was not considered for backchannel.

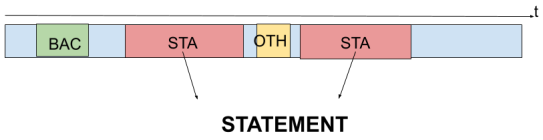

Figure 1: Each turn, following the provided segmentation by SW, is labeled considering the majority of dialogue acts that forms the turn.

gence effects for energy and speech rate. The reduction of distance between speakers in the second half of the conversation results in both the STA+OPI and BAC+AGR, showing that energy is the variable that mostly exhibits convergence. These results compared to the one of Table 10 could be partially explained by the fact that the performances of the tagger are better in this case. Increasing the precision reduces the noise of mislabeled turns that influence the filtering of DA.

6 Discussion

In this study we scrutinized *convergence* during the course of a conversation and in a real world setting (Switchboard corpus). Our work is a step toward using conversational corpus in more controlled studies on specific phenomena by using dialogue acts as a speech activity filter that reduces noise of real conversational dialogues and at the same time study convergence that happens within conversations (most of previous work establishes convergence by comparing speakers' distance in a conversation to other conversations and not what is happening in the time course of a conversation).
Our experiment consists in comparing the speakers difference between average values of a given variable in the first and second half per conversation, for the entire corpus and for the subsets made by DA annotations.

Our results show that speakers tend to have a more similar average values in the second half of the conversation for E and R. This confirms the results obtained on experimental lab speech and it is compatible with the results at corpus level.
The second idea we developed in this study is that dialog acts can have different behaviors in regard of convergence. We then split the whole datasets into sub set corresponding to specific frequent dialogue acts. Also in this case every significant or nearly significant difference correspond to a reduction of the distance between the speakers in the second part.

This result is interesting considering that the magnitude of the effect is still present even if in the dialog acts subsets the number of data to compute average values dramatically decreases (Table 1). The same trend is also found using a DA tagger of the turns produced by the speakers. Also in this case even if the automatic tagging introduce some noise the averages distance in second half of conversation decreases as shown in 11 and 10.

Our results complement and strengthen the picture provided by the literature. They also open up the possibility to a range of studies on large corpora, including new studies taking advantage on large corpora partially controlled *a posteriori* thanks to dialogue act tagging.

As for future work, we would like to confirm our results by replicating them on the larger Fisher corpus (Cieri et al., 2004). We also plan to tag the entire corpus in order to increase the statistical power as done for Switchboard. Finally, we would like to articulate out findings with the result on more local interpersonal dynamics such as *synchrony* (Levitan and Hirschberg, 2011).

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.

Jeremy Auguste, Robin Perrotin, and Alexis Nasr. 2018. Annotation en actes de dialogue pour les conversations d'assistance en ligne. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 577.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23.

Francesca Bonin, Céline De Looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli, and Nick Campbell. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. In *Proceedings of 14th Annual Conference of the International Speech Communication Association*, Lyon, France.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus:

a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

U Cohen Priva and C Sanker. 2018. Distinct behaviors in convergence across measures. In *Proceedings of the 40th annual conference of the cognitive science society. Austin, TX*.

Uriel Cohen Priva, Lee Edelist, and Emily Gleason. 2017. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, 141(5):2989–2996.

Jens Edlund, Mattias Heldner, and Julia Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *Tenth Annual Conference of the International Speech Communication Association*.

Florian Eyben and Björn Schuller. 2015. opensmile:): the munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, 6(4):4–13.

Simone Fuscone, Benoit Favre, and Laurent Prevot. 2018. Replicating speech rate convergence experiments on the switchboard corpus. In *Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*.

H. Giles, N. Coupland, and J. Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, Studies in emotion and social interaction, pages 1–68. Cambridge University Press.

Howard Giles and Peter Powesland. 1997. Accommodation theory. In *Sociolinguistics*, pages 232–239. Springer.

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*, volume 1, pages 517–520. IEEE.

A Kuznetsova, P Bruun Brockhoff, and R Haubo Bojesen Christensen. 2014. lmertest: tests for random and fixed effects for linear mixed effects models. *See https://CRAN. R-project. org/package= lmerTest*.

Rivka Levitan. 2014. *Acoustic-prosodic entrainment in human-human and human-computer dialogue*. Ph.D. thesis, Columbia University.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Uwe D Reichel, Katalin Mády, and Jennifer Cole. 2018. Prosodic entrainment in dialog acts. *arXiv preprint arXiv:1810.12646*.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Richard L. Street. 1984. Speech convergence and speech evaluation in fact-finding iinterviews. *Human Communication Research*, 11(2):139–169.

Khiet P Truong and Dirk Heylen. 2012. Measuring prosodic alignment in cooperative task-based conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.