Filtering conversations by dialogue act labels for improving corpus-based convergence studies

Simone Fuscone[1,2] and **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

During an interaction the tendency of speakers to change their speech production to make it more similar to their interlocutor's speech is called *convergence*. Convergence had been studied due to its relevance for cognitive models of communication as well as for dialogue system adaptation to the user. Convergence effects have been established on controlled data sets while tracking its dynamics on generic corpora has provided positive but more contrasted outcomes. We propose to enrich large conversational corpora with dialogue acts information and to use these acts as filters to create subsets of homogeneous conversational activity. Those subsets allow a more precise comparison between speakers' speech variables. We compare convergence on acoustic variables (Energy, Pitch and Speech Rate) measured on raw data sets, with human and automatically data sets labelled with dialog acts type. We found that such filtering helps in observing convergence suggesting that future studies should consider such high level dialogue activity types and the related NLP techniques as important tools for analyzing conversational interpersonal dynamics.

1 Introduction

The way participants engaged in a conversation speak tends to vary depending on their interlocutor's speech. The tendency to co-adjust speaking styles in response to the partner speaking style is known as *convergence*. Convergence is presented in general and influential models of communication such as *accommodation theory* (Giles et al., 1991) or *interactive alignment* (Pickering and Garrod, 2004). This variation due to the other party has been studied for many levels of speech and language, for example in phonology (Street, 1984; Pardo, 2006) or in prosody (Levitan and Hirschberg, 2011; Truong and Heylen, 2012; Bonin et al., 2013).

Our approach aims at deepening and generalizing the investigation of convergence and related effects in real-life corpora. Considered from the angle of speech and linguistic variables, an essential aspect of conversations is their extreme variability. This is due to a large extent to different conversational activities speakers can participate in. For instance, they can enter in a storytelling sequence in which one interlocutor starts to produce mostly back-channels (Yngve, 1970) while the main speaker develops lengthy monologues. This variability makes comparison of values across participants very problematic. We propose to create subsets of similar **dialogues acts (DA)** (e.g. 'statements' and 'back-channels', see Table 1 for an illustration), resulting in homogeneous data used as a proxy to characterize the conversational activity of a given turn. We intend to create subsets consisting of turns belonging to a specific function using current Dialogue Act tagging Techniques.

Our work concerns more specifically acoustic convergence. Our definition of *convergence* comes from several studies (Edlund et al., 2009; Truong and Heylen, 2012; Cohen Priva et al., 2017), and consists of comparing distance between speakers in different parts of a conversation (See Section 3). We do not claim it is the best measure to approach inter-personal dynamics (See (Priva and Sanker, 2019)) but it is an interesting way to assess convergence within a conversation and it allows to test whether our DA based approach can help this domain.

2 Related work

Convergence has been approached at different granularity levels and for a large range of variables. In terms of granularity, studies can be *Inter-conversation* comparisons or *Intra-conversation*

, pages 148 - 153

(focusing on the dynamics within conversations). Inter-conversation comparisons range from simple inter-speaker correlation studies (Edlund et al., 2009) or, when the data allows, comparison between speech values of a speaker and his conversational partners vs. a speaker and all other non-partner corpus participants (Cohen Priva et al., 2017). Intra-conversation studies vary a lot in terms of approaches ranging from "difference-in-difference" *convergence* (Edlund et al., 2009), (Truong and Heylen, 2012), (Cohen Priva and Sanker, 2018) approaches consisting of comparing distances between speakers in different intervals to finer grained *synchrony* methods typically using sliding windows 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 picture, with a relative fragility of the effects (Fuscone et al., 2018) and raised methodological comments (See (Truong and Heylen, 2012) and (Cohen Priva and Sanker, 2018)). More precisely, for *intra-conversation* studies, (Edlund et al., 2009) found that participants tend to be more similar (in terms of gaps and pauses duration) to their partners than chance would predict. However, the absence of significant results in comparing the inter-speaker distance in the first and second halves of the conversations makes the authors conclude that convergence cannot be captured with such an approach. (Truong and Heylen, 2012) conducted a similar experiment (on intensity and pitch) on English MapTask ((Anderson et al., 1991)) with partial positive results. The dynamic nature of the phenomenon as well as the social factors render such studies difficult to be performed. These two studies were grounded on conversational corpora that are sizeable but not huge (6 x 20 minutes for the (Edlund et al., 2009); and about 60 MapTasks dialogues for (Truong and Heylen, 2012)). (Cohen Priva and Sanker, 2018) used the much bigger Switchboard corpus but use only an inter-conversation approach.

Our hypothesis is that automatic *entrainment* and strategic *adaptation* are blending in to produce *convergence* and *synchrony* phenomena. Our hypothesis is that low-level variables (such as intensity) are be more directly related to automatic entrainment, while higher-level variables (such as lexical or syntactic choices) are more prone to strategic adaptation. This could explain why more and firmer results seem to be obtained on low-level variables (Natale, 1975; Levitan, 2014).

To summarize, convergence dynamic can be difficult to track in real conversations. Our approach combines three ingredients that, to our best knowledge, were not yet brought together. First, we consider that a major reason for this difficulty comes from the heterogeneity of speech behaviors within the time-frame of a conversation. We propose to use DA to filter conversational activities from large corpora. Second, to account for *strategic adaptation* one must take precise care of speaker profiles. Our approach therefore focuses on relatively low level variables to avoid as much as possible the "adaptation" part of the interpersonal dynamics. Third, similarly to (Cohen Priva et al., 2017) our approach is based on a large conversational corpus with the intention of overcoming noise and effect small magnitude by increasing the amount of data considered.

3 Methodology

3.1 Convergence

Following (Edlund et al., 2009) and (Truong and Heylen, 2012) we divide each conversation into two halves and compare the distance between the average values of the target variables of each speaker. This provided us two values (first and second interval) for each variable and each conversation: $\Delta\overline{V}_i = \mid \overline{V}_{Ai} - \overline{V}_{Bi} \mid$, 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 corresponds to **Energy (E)**, **Pitch (F0)** and **Speech rate (SR)**. Our aim is to test the hypothesis that convergence occurs during the interaction. We therefore computed the distance 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 of 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

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

3.2 Feature processing

E and F0 are computed from the audio files with *openSMILE* audio analysis tool (Eyben and Schuller, 2015) while SR is computed using time aligned transcripts.

Energy (E): One of the issues of telephonic conversation is the distance mouth-microphone that affects measured values of voice intensity can be different across speakers and even for the same speaker across conversations. So to reduce this effect we introduce a normalization factor by 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 (as provided by Switchboard transcripts) the average E.

Pitch (F0): We computed the average in each conversational unit for each speaker.

Speech Rate (SR): We used the approach proposed by Cohen-Priva (Cohen Priva et al., 2017) that defines SR 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 SR more reliable we consider only utterances having more than 5 tokens.

4 Dialogue Act Filtering and Data Sets

Switchboard (SWBD) (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*.

642 Switchboard conversations have been segmented and annotated for DA that we will call the NXT data set (Calhoun et al., 2010). [2] The DA-tagged set has been simplified to 42 tags but a few of them are dominating the distribution (Statement: 36%, Acknowledgment: 19%, Opinion: 13%), illustrated in Table 1. See (Stolcke et al., 1998) for details.

DA type	Example
Statement	*" that was pretty heartrending for her"*
Opinion	*"money seems to be too big of an issue."*
Backchannel	*"Uh-huh."*
Agree.	*"you're right"*

Table 1: Examples for the DA types used.

Automatically tagged data set We create a **turn tagger**, using 3 categories, corresponding to *Statement+Opinion* (STA+OPI), *Backchannel+Agreement* (BAC+AGR) and *Other* (OTH) which includes all the other DA. This grouping was obtained by first considering only the DA dominating the distribution. Then we manually checked many examples of each DA and figure out that although functionally different *statements* and *opinions* on the hand; and *backchannel* and *Agreement* those group were similar enough for our current purposes. The former has a *main speaker* nature while the later have a much more *listener* nature (see Table 1).

We used as train, development and test set the NXT Switchboard corpus that contains annotated DA for 642 conversations. Since the DA segmentation does not match the turn segmentation, we label each turn of the corpus by assigning the majority class, among the DA tags used in the turn. The resulting distribution is 52% STA+OPI, 25% BAC+AGR and 23% of OTH. 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 levels 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 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 (LSTM). We used 80% of Switchboard data as training set, 10% for development and 10% for the test set. The F1 score

[1] Speakers therefore do not know each other.

[2] We use this version of DA as it contains alignment to the transcripts, contrarily to the SWDA bigger data set (Jurafsky et al., 1997).

of the DA tagger is 81% on the test set, the details for each category is reported in table 2. The F1 score of the class OTH is low compared to the other 2 classes.

Class	Precision	Recall	F1
Bc+Agr.	0.88	0.85	0.86
St.+Opi.	0.84	0.92	0.87
Oth.	0.62	0.49	0.55

Table 2: Prediction score of our turn tagger.

5 Results

Our question is whether we can observe more reliably interpersonal dynamics in raw, manually DA-tagged (small) or automatically DA-tagged (large) data sets. An underlying question is whether the noise introduced by the DA-tagging uncertainty and / or the data size reduction is compensated by the gain in homogeneity between the material that is compared.

5.1 DA-tagging contribution

We first report the results in the case of the whole data set without DA (SWBD) and manually DA-tagged (NXT). The results are summarized in Table 3.

	All	St.	Opi.	Bc.
SWBD (180h)	**E** - R	X	X	X
NXT	E - X (41h)	E - **R** (17h)	- - - (7h)	E - - (1h)

Table 3: Manual tagging results summary (E: Energy; P: Pitch ; R: Speech Rate; - : no significance; normal font : p-value$\leq$0.05 ; **bold** : p-value $\leq$ 0.01). See Table 5 in Appendix for details.

When differences are significant, it is always in the direction of reduction of the distance (See Appendix for the details). We observe that concerning *Statement*, with less than 10% of the original data, the method allows one to observe the same effect as in the whole Switchboard (and reaches a higher level of significance for SR). The *Statement* subset shows convergence for E and SR. *Statement*-filter seems to homogenize the data set by filtering out particular back-channels and strong disfluencies (type *abandoned*). This helps observing the effect for SR. Contrarily, the wide variety of *statements* in terms of utterance duration could be an issue for F0 since contours and physiological-related decreasing slope could result in a lot of noise for this variable. There are no positive results on *Opinion* maybe due to larger variability or consistency in this label. *Back-channel* although keeps the effect on the E but, due the nature of this speech act, SR is not relevant. F0 doesn't show any significant results. This probably can be explained considering that F0 is a more complex variable and the average approach is not able to capture more subtle characteristics of F0 (Reichel et al., 2018).

5.2 Automatic Tagging results

As explained above, in the experiment on automating tagging we merged the most similar frequent DA. The automatically tagged corpus preserved the results from the raw data sets. Similarly for the manual version, automatic tags filtering helped reaching better significance for SR on *Statement+Opinion* utterances as summed-up in Table 4. Back-channels were excluded from the SR experiment since our measure of SR isn't reliable on such short utterances.

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

Table 4: Automatic tagging results summary (E: Energy; P: Pitch ; R: Speech Rate; - : no significance; normal font : p-value$\leq$0.05 ; **bold** : p-value $\leq$0.01). See Table 6 in Appendix for details.

6 Discussion

We scrutinized *convergence* during the course of a conversation and in a real world setting (Switchboard corpus). The positive results in our experiments complement the picture provided by the literature by showing that convergence effects do happen in the time course of conversation of generic corpus. Moreover, we open up the possibility of a range of new studies taking advantage on arbitrary large corpora partially controlled *a posteriori* thanks to automatic dialogue act tagging.

Acknowledgments

This project received funding from the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No713750. Also, it has been carried out with the financial support of the Regional Council of Provence-Alpes-Côte d'Azur and with the financial support of A*MIDEX (ANR-11-IDEX-0001-02) and ILCB (ANR-16-CONV-0002).

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.

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.

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.

D Jurafsky, E Shriberg, and D Biasca. 1997. Switchboard dialog act corpus. *International Computer Science Inst. Berkeley CA, Tech. Rep*.

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

Uriel Cohen Priva and Chelsea Sanker. 2019. Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1).

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.

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

A Appendices

SWBD	*Entire Corpus (180 hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.063**	**0.012**	$\mathbf{7 \times 10^{-7}}$
F0-Mean	-0.044	0.021	0.490
SR-Mean	**-0.049**	**0.024**	**0.046**
NXT	*Whole DA-tagged subset (41 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.054**	**0.021**	**0.026**
F0-Mean	-0.057	0.040	0.158
SR-Mean	**-0.106**	**0.047**	**0.026**
NXT	*Backchannel Tag Subset (1 Hour)*		
Feature	Estimate	std	p-values
E-Mean	**-0.082**	**0.041**	**0.045**
F0-Mean	0.043	0.022	0.491
NXT	*Statement Tag Subset (17 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.071**	**0.023**	**0.032**
F0-Mean	-0.025	0.038	0.653
SR-Mean	**-0.123**	**0.049**	**0.012**
NXT	*Opinion Tag Subset (7 Hours)*		
Feature	Estimate	std	p-values
E-Mean	-0.061	0.033	0.627
F0-Mean	-0.032	0.053	0.552
SR-Mean	-0.096	0.061	0.115

Table 5: 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.

Auto	*Energy*		
CLASS	Estimate	std	p-values
STA+OPI	**-0.055**	**0.011**	$4 \cdot 10^{-6}$
BAC+AGR	**-0.079**	**0.028**	**0.006**
Auto	*Pitch*		
CLASS	Estimate	std	p-values
STA+OPI	-0.035	0.038	0.353
BAC+AGR	0.053	0.028	0.192
Auto	*Speech rate*		
CLASS	Estimate	std	p-values
STA+OPI	**-0.075**	**0.021**	**0.008**

Table 6: Parameters of our linear model for Energy, Pitch and Speech Rate for the corpus automatically tagged. Speech Rate was not considered for back-channels.