Using BERT to measure coordination in bi-lingual dialogue: An information-theoretic approach

Bill Noble and **Fahima Ayub Khan**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{bill.noble@, fahima.ayub.khan@}.gu.se

Abstract

In this work, we propose an information theoretic metric of coordination in dialogue, namely the information gain afforded to a language model supplied with dialogue context, versus the same model without dialogue context. We experiment with this metric using, multilingual BERT to test the hypothesis that code-switching is a coordination mechanism in multilingual dialogue.

1 Introduction

Face-to-face dialogue requires that participants coordinate on multiple levels of interaction. Coordination is both *procedural* (coordinating on the structure of the interaction) and semantic (coordinating to establish *what has been said*). While these two aspects of interaction are difficult to peel apart, it is clear that most work attempting to quantify coordination is chiefly concerned with procedural coordination. Most statistical analyses of content in this area have sought to measure *alignment*, a related but distinct phenomenon. Face-to-face dialogue is a complex activity that requires coordination on multiple levels. Coordination is both *procedural*, i.e., coordinating on navigating the different stages of a joint project (Mills, 2014; Bangerter et al., 2020) and semantic, that is, referential convergence (coordinating to establish *what has been said*). While these two aspects of interaction are difficult to peel apart, it is clear that most work attempting to quantify coordination is chiefly concerned with referential coordination. In parallel, most statistical analyses of content in this area have sought to measure *alignment*, a related but distinct phenomenon.

The measure of coordination proposed in this paper operates on the intuition that coordination in diolague can be characterized as involving interloquitors who follow the maxims set out below in the manner of Grice's maxims:[1]

1. Act in such a way that *makes use of* the dialogue context.
2. Act in such a way that *creates contexts* that are useful to your interlocutors.

In this work, we propose the *information gain* afforded by dialogue context as a measure of dialogical coordination. As a proof of concept, we use this this metric with multi-lingual BERT to test the hypothesis that code-switching is a mechanism for achieving coordination in multi-lingual dialogue.

Before embarking on the computational aspects of this work, we will give some background on the two main dialogue concepts under consideration, coordination and code-switching.

2 Background

Coordination in dialogue Turn-taking is a classic example of procedural coordination. Speakers coordinate to minimize gaps between turns while at the same time keeping speech overlaps brief (Sacks et al., 1974). This is achieved through a complex system of cues that include eye gaze (Sekicki and Staudte, 2018), head movement (Maynard, 1990), and incremental/predictive processing of utterances (Schlangen and Skantze).

Most work in this area that considers the content of utterances focuses on *alignment*, often as a proxy for coordination. Certain psycholinguistic models of language processing even posit alignment as the basis of successful interaction (Pickering and Garrod, 2004). However, coordination cannot be reduced to priming-driven alignment, since true coordination in dialogue often requires superficially divergent behavior among participants

[1] While these maxims *characterize* the behavior of coordinative speakers, we do not mean to suggest that it is *from* these maxims that real-life speakers act. The cognitive mechanisms that bring about coordination are a separate question.

(Healey et al., 2014). Consider the following exchange:

A: Would you like some tea?
B: Yes, thank you. (1)

We understand speaker B's utterance as demonstrative of coordination since it only makes sense in the context of what A said — it *makes use of* the dialogue context.

Consider, on the other hand:

A: Would you like some tea?
B: I would like some tea. (2)

While perfectly acceptable as a stand-alone sentence, (2) is a little odd in conversation. This utterance does not make use of the dialogue context afforded by A. Note that this is true in spite of the fact that most measures of alignment would rate (2) as exhibiting greater alignment with A than (1).

Notable exceptions to the alignment-centered approach have focused on the coordinative role of specific dialogical phenomena, including interjective feedback or *backchannels* (Healey et al., 2018; Howes and Eshghi, 2019), clarification requests (Healey et al., 2011), repair (Ginzburg et al., 2007; Purver et al., 2018), and disfluencies (Ginzburg et al., 2014). While the present study does not focus on these phenomena, future work could explore the utility of our methodology in their analysis.

Code-switching Code-switching in interaction has been widely studied within conversation analytic (CA) approaches using spontaneous bilingual speech, however, by design, such approaches preclude experimental testing to understand the communicative effect of code-switching in dialogue under controlled conditions. Nonetheless, insights from cross-disciplinary studies on the interactive features of code-switching contribute to formulating testable hypotheses on bilingual dialogue. The earliest CA approaches to bilingual speech examined the sequential use of different languages in the same conversation. It has been established that bilingual speakers are more likely to code-switch with members of their language community, especially in highly bilingual environments.

The discourse functional properties of different types of code-switching (such as *alternational* and *insertional* code-switching) have been studied in experimental and corpus-based approaches. Mackey (2000) observed that alternational code-switching is governed by changes in topic and the people involved in the conversation. Similarly, Wei and Milroy (1995) note that switching to a language different from that of the previous utterance signals an imminent dispreferred response, mirroring the contrast in the flow of the content. Alternational code-switching is one of the three distinct intra-sentential switching patterns defined by Muysken (2000) where switching happens between utterances or at clause boundaries. On the other hand, insertional switching refers to instances when a word or a constituent from one language is inserted into the syntactic frame of another language. A study by Myslín and Levy (2015) investigated the discourse function of insertional switches at the end of Czech-English code-switched utterances. They noted that more informative utterance endings were marked by switching to another language.

In this work, we do not distinguish between insertional and alternational codeswitching, since this would require detailed syntactic annotation, but we do distinguish between *within-utterance* and *between-utterance* code-switches.

3 Method

An autoregressive language model makes next-token predictions by considering the tokens that come to the left of the target position. In this way, the entropy of such a model's prediction distribution is an estimate of how (un)predictable the next token is. In the following, we use the term *dialogue model* loosely to refer to language models that also consider the left context of a dialogue, in addition to that of the current utterance. We will first introduce the proposed dialogue model-based coordination metric before moving on to describe how we use BERT as a rudimentary dialogue model.

3.1 Information gain of the dialogue context

In the following, let M be a dialogue model; that is, given a tokenized utterance $u = (t_1, ..., t_n)$ and dialogue context c, M gives us two functions:

$$\mathrm{M}^u(u, i)(x) = p_{\mathrm{M}}(x \mid t_1, ..., t_{i-1}), \qquad (1)$$

and

$$\mathrm{M}^d(c, u, i)(x) = p_{\mathrm{M}}(x \mid c; t_1, ..., t_{i-1}), \qquad (2)$$

such that both M^u and M^d give a prediction of the next token, x (i.e., a probability distribution over the vocabulary). The difference between the two functions (and the reason we regard M as a *dialogue model*) is that M^d incorporates information provided by some notion of dialogue context.

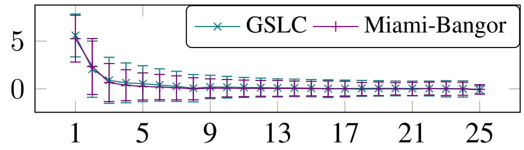

Figure 1: Mean un-adjusted contextual information gain by token position. Error bars show standard deviation. Information gain of 0 means that the context provides no additional information for guessing the masked token.

With this in hand, we can estimate the token-level information gain afforded by the dialogue context, the *contextual information gain*, with respect to the dialogue model M as follows:

$$\mathsf{IG}_{\mathrm{M}}(u, i) = H(\mathrm{M}(u, i)) - H(\mathrm{M}(u, i, c)), \quad (3)$$

where H is the entropy of a discrete probability distribution, as it is usually defined.

Like language model perplexity, this metric suffers from the difficulty that it varies a lot depending on the position of the token — dialogue context is much more helpful predicting early tokens since there is less else to go on (see Figure 1). For this reason, we compute and μ_i, σ_i, the mean and population standard deviation of $\mathsf{IG}_{\mathrm{M}}(u, i)$ over all utterances for each position and define the adjusted metric as follows:[2]

$$\mathsf{IG}^{*}_{\mathrm{M}}(u, i) = \frac{\mu_i - \mathsf{IG}_{\mathrm{M}}(u, i)}{\sigma_i} \quad (4)$$

Intuitively, this metric tells us, for a given utterance, how much more helpful than average the dialogue context is in the task of predicting token i.

3.2 Multi-lingual BERT

Multi-lingual BERT (M-BERT), introduced by Devlin et al. (2019), is a transformer-based language model with the same architecture as the standard BERT model, but trained on 104 corpora of text from different languages (mostly Wikipedia). Although m-BERT was trained on primarily monolingual texts, Pires et al. (2019) finds that it is able to perform tasks in code-switched environments. The pre-trained English language BERT has been used for dialogue understanding with some success. The current study will serve as another test of BERT's ability to generalize to dialogue contexts, although we leave fine-tuning for future work.

In pre-training, M-BERT is shown two sentences at a time, separated by a `[SEP]` token. The two sentences are consecutive in their source corpora 50% of the time, so M-BERT learns to improve its token predictions based on context beyond the sentence level. We use M-BERT as a rudimentary dialogue model by providing it with adjacent utterances. For a pair of adjacent utterances u_1 and $u_2 = (t_1, ..., t_n)$, we define $\mathrm{BERT}^d(c, u_2, i)$ as the predictions that M-BERT gives for position i with u_1 and u_2 as input and tokens $t_i, ..., t_n$ masked out. $\mathrm{BERT}^d(u_2, i)$ is computed in the same way, but with all of u_1 masked out (see Figure 2).

4 Data

	GSLC	Miami-Bangor
context	288	5 997
target	288	5 992
both	27	1 421
between	148	9 489
any	605	15 023
total pairs	111 043	39 043

Table 1: Number of code-switched utterance pairs by code-switch type in each corpus. *Context* and *target* switches are similar in number, since most target utterances appear as the context for the following utterance.

In the experiments, we use data from two dialogue corpora: the Bangor-Miami corpus, and the Gothenburg Spoken Language Corpus (GSLC).

The Bangor-Miami corpus (Deuchar, 2010) consists of 41 informal, mostly dyadic, face-to-face conversations between bilingual speakers of English and Spanish.[3] The corpus was transcribed in the CHAT format (MacWhinney, 2022), and annotated for language at the token level.

The GSLC consists of 360 dialogues in a mix of dyadic, multi-party and one-party dominant scenarios. The conversations range various genres, including discussions, interviews, informal conversation, and task-oriented dialogue. The dialogues are transcribed using the Göteborg Transcription Standard (Nivre et al.) and use Modified Standard Orthography (Nivre, 1999) for Swedish orthographic transcription. Non-Swedish tokens are annotated with language of origin. Of 1121 non-Swedish tokens in the GSLC, the majority (737) are annotated as English, and the remainder are mainly Esperanto

[2] This is done separately for each corpus.

[3] We excluded the `Maria` section dialogues, where the partner utterances were redacted. The original corpus had 56 conversations.

	0	1	2	3	4	5	6	7	8	9
unmasked text:	[CLS]	how	are	you	[SEP]	good	how	are	you	[SEP]
unconditioned input:	[CLS]	[MASK]	[MASK]	[MASK]	[SEP]	good	how	[MASK]	[MASK]	[SEP]
contextualized input:	[CLS]	how	are	you	[SEP]	good	how	[MASK]	[MASK]	[SEP]

Figure 2: Example inputs used for measuring M-BERT's token prediction entropy, $\mathrm{BERT}(x \mid t_1, ...t_{i-1})$ and context-conditioned entropy, $\mathrm{BERT}(x \mid u_1; t_1, ...t_{i-1})$.

(243) and Norwegian and/or Danish (82).[4]

Both corpora were preprocessed to remove disfluency markers, overlaps, and other annotation symbols that do not have conventional textual representations. Orthographic transcriptions were normalized to their standard textual forms.

5 Experiments

We measure $\mathsf{IG}^*_{\mathrm{BERT}}(u, i)$ for every (non-dialogue-initial) utterance in both corpora. We first tokenize the utterances with a wordpiece tokenizer[5] and limit the length of utterances to 25 tokens.[6]

We consider code-switching at the token level and look at between-utterance and within-utterance code-switching separately, since there is some evidence that they play different coordinative roles. In each of these scenarios, we compare the contextual information gain when there is a code-switch in the context (i.e., the previous utterance) and when there is not. We also consider whether the token in question itself switches from the language of the previous token (or utterance).

CS in context	N		Y	
CS b/t utterances	N	Y	N	Y
GSLC	0.00	-0.05	0.04	0.19
Miami-Bangor	-0.05	0.22	-0.04	0.12

Table 2: Mean $\mathsf{IG}_{\mathrm{BERT}}(u, 1)$ (for utterance-initial tokens), stratified by (1) whether the context (previous utterance) includes a code switch and (2) whether the initial token switches from the language that ended the previous utterance (between-utterance switching).

For utterance-initial tokens (Table 3), we measured higher than average when both the target token is a code-switch *and* there is a code-switch in the previous utterance, for both corpora. We also measure high contextual information gain when there is no code-switch in the previous utterance, but only for the Miami-Bangor corpus. This could have to do with the fact that code-switching is a much more integrated aspect of the communicative practice of that community (as evidenced by its relative prevalence). Thus between-utterance code-switching is predictable from non-switched contexts, whereas in the GSLC, non-code-switched contexts are not such that they predict following utterances in a different language.

CS in context	N		Y	
CS target token	N	Y	N	Y
GSLC	0.00	-0.22	0.09	0.04
Miami-Bangor	0.01	-0.21	-0.04	-0.17

Table 3: $\mathsf{IG}^*_{\mathrm{BERT}}(u, i)$ for non-utterance-initial tokens ($i > 1$), stratified by (1) whether the context (previous utterance) includes a codeswitch, and (2) whether the token switches from the language of the previous token (within-utterance switching).

In the case of non-initial tokens (Figure 2), information gain is again lower for switched tokens when there is no switch in the context; that is, contexts with switches are better at predicting switch tokens than contexts without. Here, however, we do not see the above-average information gains that we saw for between-utterance switching. It is worth noting, however that as these measures are normalized, a negative score does *not* mean negative information gain — in fact, while the variance is high, the information gain afforded by dialogue context is nearly always positive in aggregate.

In the two experiments we see differences between the corpora. This suggests that the coordinative function of code-switching may depend on the frequency and type of switching that is used. We also see that IG^* is clearly measuring something very different from alignment. Consider for example, the relatively high information gain measured for between-utterance switching in the MIami Bangor corpus when there is no switching in

[4] All of the Esperanto tokens come from a single dialogue, A7925012.

[5] We use the M-BERT and corresponding tokenizer implementation and model weights provided by the Transformers Python library, version 4.12.2 (Wolf et al., 2020).

[6] In cases where an utterance is longer than 25 tokens (11.4% and 0.7% of and utterances in GSLC and Miami-Bangor, respectively), we truncate context utterances to the final 25 tokens and target utterances to the initial 25 tokens.

the context.

6 Conclusion

In this abstract, we have proposed a new information-theoretic metric for measuring one aspect of coordination in dialogue, namely, to what degree the dialogue context can be used to incrementally predict the next utterance. We used this metric in a preliminary investigation of the hypothesis that code-switching has a coordinative function in multi-lingual dialogue. We found that there are differences in contextual information gain depending on whether the target token is a switch from the language of the previous token and depending on whether there is a code-switch in the context.

The results reported here are preliminary. What constitutes coordination depends on many factors beyond the content of what is said, including conversational genre, community norms, and relationships between interlocutors. The utility of measure $\mathsf{IG}^*_{\mathrm{M}}$ as a measure of coordination entirely dependent on how well M models these different aspects of context. No doubt M-BERT, which was trained on non-dialogical text, leaves a lot to be desired. In future work, we will experiment with more sophisticated dialogue models and compare the contextual information gain measured by different models. We will also explore how contextual information gain relates to other dialogue phenomena related to coordination.

References

Adrian Bangerter, Eric Mayor, and Dominique Knutsen. 2020. Lexical entrainment without conceptual pacts? revisiting the matching task. *Journal of Memory and Language*, 114:104129.

Margaret Deuchar. 2010. BilingBank Spanish-English Bangor Miami Corpus.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Ginzburg, Raquel Fernandez, and David Schlangen. 2007. Unifying Self- and Other-Repair. In *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue,*, page 7.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(0):9–64.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science*, 10(2):367–388.

Patrick GT Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, pages 11–13. Citeseer.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

Christine Howes and Arash Eshghi. 2019. Interjection as coordination device: feedback relevance spaces. In *16th International Pragmatics Conference*, Hong Kong.

William F Mackey. 2000. The description of bilingualism. *The bilingualism reader*, pages 26–54.

Brian MacWhinney. 2022. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition.

Senko K. Maynard. 1990. Conversation management in contrast: Listener response in Japanese and American English. *Journal of Pragmatics*, 14(3):397–412.

Gregory J Mills. 2014. Dialogue in joint activity: Complementarity, convergence and conventionalization. *New ideas in psychology*, 32:158–173.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, pages 871–905.

Joakim Nivre. 1999. Modifierad Standardorthografi (MS06). Technical Report Version 6, Institutionen för lingvistik, Göteborgs universitet.

Joakim Nivre, Jens Allwood, Leif Grönqvist, Magnus Gunnarsson, Elisabeth Ahlsén, Hans Vappula, Johan Hagman, Staffan Larsson, Sylvana Sofkova, and Cajsa Ottesjö. Göteborg Transcription Standard. Technical Report Version 6.4.

Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–225.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational Models of Miscommunication Phenomena. *Topics in Cognitive Science*, 10(2):425–451.

H. Sacks, E. Schegloff, and G. Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. 50(4):40.

David Schlangen and Gabriel Skantze. A General, Abstract Model of Incremental Dialogue Processing. page 9.

Mirjana Sekicki and Maria Staudte. 2018. Eye'll Help You Out! How the Gaze Cue Reduces the Cognitive Load Required for Reference Processing. *Cognitive Science*, 42(8):2418–2458.

Li Wei and Lesley Milroy. 1995. Conversational code-switching in a chinese community in britain: A sequential analysis. *Journal of Pragmatics*, 23(3):281–299.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.