Interaction with a Computer Increases Human Linguistic Alignment

Authors: Simon A. M. Hansen & Line Kruse
GitHub: https://github.com/simonamhansen/HumanComputerAlignment-

Abstract

While linguistic alignment has commonly been understood as an automatic priming process, recent advances propose that it should be conceptualized as an emerging dynamic property, which is highly context-dependent and to some extent strategic. The current paper investigates how three levels of alignment varies as an effect of three types of communication partners (Human-Human, Human-Wizard of Oz, and Human-Computer). The results suggest that humans align more when interacting with computers compared to another human. Further, while alignment decreases over time in H-H interactions, it seems to increase in H-C interactions. Potential confounding effects of the data employed are discussed.

Introduction (Line)

As technological development advances humans increasingly interact with computer systems both physically and verbally. Subsequently, it becomes more and more relevant to investigate the particular nature of such interactions, both in terms of understanding how we design successful computer systems for particular tasks, but also to assess how such interactions may depend on, as well as influence, our own cognitive and linguistic processes. Do we represent such systems as social agents? And if so, does this have an influence on natural human processes of social interactions? The current study aimed to address these issues in relation to linguistic alignment in social interactions.

The nature of linguistic alignment

Linguistic alignment is defined as the tendency for two interlocutors to converge on similar linguistic expressions (Duran et al., 2019). Alignment functions to coordinate the situational models of two interlocutors in terms of production as well as comprehension, which is thought to be essential to successful communication (Koulouri, Lauria & Macredie, 2016). One of the most prominent models of alignment, the interactive alignment model (Pickering & Garrod, 2014), understands alignment as occurring through automatic priming processes. That is, encountered linguistic structures prime the subsequent use of similar structures in an

unconscious and automatic fashion. Further, alignment at one linguistic level (e.g., lexical) is thought to promote alignment at other levels (e.g., syntactic) (Branigan & Pearson, 2006). However, accumulating evidence indicates that this mimicry behaviour is highly dependent on the particular context. A recent approach conceptualizes alignment as a synergy; an emergent dynamic property of interaction that cannot be reduced to individual linguistic processes, but emerges and adapts according to the constraints and goals of the given situation (Fusaroli, Rączaszek-Leonardi & Tylén, 2013). The crucial difference here is the notion that indiscriminate alignment may not always be beneficial to the interaction. Rather, as the interactional process stabilizes over time, less negotiation and structuring of representation might be needed leading to a decrease in alignment. In fact, evidence indicates that alignment decreases over time in human linguistic interactions, suggesting that it may not be an entirely automatic processes (e.g. Wang, Reiter & Yen, 2014).

Alignment in Human-Computer Interactions (HCI)

As linguistic alignment serves to align mental situational models it implies the perception of a communication partner as a thinking agent. However, it has commonly been found that humans align with computer systems to a similar extent as when interacting with another human. People adjust their own behaviour both at a non-linguistic level, such as speaking rate, amplitude, pause frequency (Koulouri et al., 2016; Bell, Gustafson, & Heldner, 2003), as well as on linguistic expressions including lexical and syntactic structures (Branigan, Pickering, Pearson, McLean & Brown, 2011; Branigan et al., 2004). In some cases, alignment has even been found to be greater in HCI compared to Human-Human Interactions (HHI) (Branigan et al., 2004). While these observations may appear surprising, they reveal something fundamental about linguistic alignment; that it is dependent on the assumptions we make about the knowledge and beliefs of the communication partner. Branigan and colleagues (2004) demonstrated that people align more with computer systems thought to be basic and simple, compared to systems perceived to be complex. Such findings have been interpreted as a way to compensate for potential errors arising from misunderstanding (Schmader & Horton, 2017). That is, by relying on expressions used by the computer, people ensure that the system understands them and hence increase the success of the interaction. Additionally, the increased alignment with systems perceived to be simple compared to systems perceived to be complex, occurred independently of the actual capabilities of the system. Thus, in contrast to HHI, people do not seem to update beliefs about the capabilities of the partner in the case of

HCI. Hence, although alignment in HCI seems to imply the perception of the computer as a social agent, the interaction with computers appear to prompt different conversational strategies in at least some aspects.

The use of conversational strategies may potentially influence our own linguistic processes and subsequent cognitive processing. Linguistic alignment, for instance, has been argued to facilitate category acquisition more efficiently than individual exploration. Schmader and Horton (2017) showed that individual category conceptualization following a conversation was modulated by the way in which referents were negotiated during dialogue. This negotiation differed in HCI and HHI, such that a narrower range of content words was used in HCI, which subsequently determined the type of conceptual categorization made by participants. Thus, linguistic interaction with a computer can have cognitive consequences extending beyond the particular conversation.

The current study

Although much work has been done on linguistic alignment in HCI, discrepancies in the literature remain regarding the extent to which alignment is automatic or strategic, as well as the extent to which the nature of alignment is similar in HCI and HHI. These may result from general methodological problems in the literature. First, interactions tend to be based on very specific tasks to be solved jointly (e.g. Maes, Marcelis & Verheyen, 2007; Branigan & Pearson, 2006; Pearson, Hu, Branigan, Pickering & Nass, 2006; Brennan, 1991). If alignment is context-dependent the task constraints are very likely to influence both the extent and nature of alignment. More naturalistic dialogue is likely to improve measures of alignment and provide clarity on its exact nature. Second, different methods are employed to calculate alignment, including word counts (Pearson et al., 2006), counts of syntactic structure (e.g., Direct Object/Indirect Object in Branigan & Pearson, 2006), and Cohen's Kappa (Pearson et al., 2006). Third, most studies investigate only one level of linguistic alignment (either lexical, syntactic or semantic). As models of alignment suggest that alignment at one level spreads upwards and prompts alignment at other levels, it may be beneficial to account for interactions between these. Lastly, most studies assess alignment of an entire conversation independently of the development over time. However, particularly the debate between automatic and strategic components of alignment may benefit from investigating the influence of time. Automatic alignment implies cumulative structural priming (Duran et al., 2019). Hence, it

would be expected that alignment continuously increases throughout the conversation, or at least remains stable. Additionally, this development should be similar in HCI and HHI, as alignment should occur whenever any linguistic structure is encountered regardless of the situation. Contrary, if alignment is mainly driven by strategic components we may expect a difference in alignment over time, depending on the need to improve smoothness of the interaction. As this need may differ in HCI and HHI as a result of the computer's capabilities and the human's beliefs about these, the development in alignment would be expected to differ between these conditions.

The current study aimed to investigate the development in alignment in three types of conversational context; Human-Human Interaction (HHI), Human-Wizard of Oz Interaction (HWI), and Human-Computer Interaction (HCI). A Wizard of Oz represents a communicative partner who is human, but is perceived by the participant to be a computer. These three conditions allowed us not only to detect alignment differences in HCI and HHI, but also whether these differences were mostly dependent on the actual capabilities of the computer system or the beliefs about these capabilities. Since the Wizard of Oz is more capable of communicating than a computer system, we might expect less need for compensation and hence less alignment than in HCI. Further, since participants believed it was a computer, we might expect more alignment compared to HHI. Alignment was calculated on a turn-by-turn basis allowing us to assess development over time. Further, alignment was measured at three linguistic levels including lexical, syntactic, and semantic alignment, in order to investigate whether they are distinctly affected by the experimental conditions. Lexical alignment refers to the reuse of particular word types. Syntactic alignment represents the extent to which the grammatical structure resembles the structure of the partner's utterance. Lastly, semantic alignment expresses convergence on the meaningful content of utterances, i.e., the extent to which two expressions refer to the same conceptual content independent of the particular lexical and syntactic choice. Three hypotheses were tested:

H1) Alignment is greater in all three conditions compared to the baseline.

H2) Humans will align more with their partner when interacting with a Wizard of Oz compared to another human, and alignment will further increase when interacting with a computer compared to a Wizard of Oz.

H3) Human alignment will exhibit distinctive development over time in the three conditions.

Data (Simon)

The current study investigated alignment in three different datasets. The first was a Human-Computer type of interaction, where random people chatted with the winning chatbot Mitsuko at the Loebner Prize 2019 competition (Worswick, 2019).

The second dataset was the Coached Conversational Preference Elicitation (CCPE) dataset (Radlinski et al., 2019). This data was collected by paying two crowd workers from an unnamed open source platform. Two crowd workers were paired, and one was assigned the role of the Wizard of Oz (WoZ), while the other was assigned the role of the user. The user would speak into the computer microphone which would be played to the WoZ. The WoZ would then write their answer and a text-to-speech program would convert this into synthesised speech, which would be played for the user. Another important point was that the assistant was provided with a manuscript and had to make the user talk about their movie preferences. In that way this dataset cannot be characterised as free dialogue. Another important point is that the authors mention that some of the users might have suspected that the WoZ was actually a human. The dataset was retrieved as a *.json* file containing different transcriber notes.

The last dataset was the The Switchboard-1 Telephone Speech Corpus which is a part of the Penn Treebank (Godfrey & Holliman, 1993). This dataset was collected by pairing two Americans together, who did not know each other. They would then have a dialogue with one another over the telephone for 5 minutes on a given topic. Data was retrieved as *.txt* files and was stripped for headings and transcriber notes. Further, regular expressions were used to remove odd characters.

For all three datasets, the utterances of both interlocutors were extracted and saved as one text file per conversation, with one row containing speaker and the other row containing the utterance. The characteristics of the three datasets are summarised below in table 1. It is clear that there are major differences in both size and characteristics of the datasets, which should be accounted for.

***Table 1**: Characteristics of the three datasets used in the analysis*

Dataset and author	N of conv.	N of Turns/conv.	Interaction type	Characteristics
Mitsuko (Worswick, 2019)	46	Mean: 25.2 SD: 22.7	Human-Computer	Written, free conversation
CCPE (Radlinski et al., 2019)	502	Mean: 17.5 SD: 5.70	Human-Wizard of Oz	Semi-spoken, preference elicitation, set topic
Switchboard-1 (Godfrey & Holliman, 1993)	1155	Mean: 37.9 SD: 20.8	Human-Human	Spoken, set topic

Methods (Simon)

Alignment calculation:

In order to obtain lexical, semantic and syntactic alignment scores the ALIGN Python library (Duran et al., 2019) was employed. This library has been designed in an attempt to create an easily applicable tool that allows for comparison of different studies. In general, the ALIGN pipeline can be broken down into two parts. The first is a pre-processing stage where the dialogue data is cleaned, and different features are extracted. The latter step is where the alignment measures are calculated. For this study all settings were kept at default.

The pre-processing step contains multiple classical NLP tasks such as spell-checking, lemmatization and stop word removal. It takes as an input one text file per conversation containing two rows. The first specifying speaker and the latter specifying the utterance produced. First, any character that is not a letter or a whitespace is removed. In addition, regular expressions are used to remove common fillers such as “uh” and “huh”. Another important element of this first step is that any utterance containing less than 2 words is removed from the data, as syntactic and lexical alignment is calculated on bigrams. You could also argue that a one-word utterance contains too little information to calculate a meaningful alignment score. If a speaker has two utterances right after each other these are merged together. This is done so that alignment is always calculated between speakers rather than how much people align to themselves. The words are then spell checked. This includes writing out common contractions (ain’t, can’t, etc.). Spell checking is carried out using a Bayesian spell-checking algorithm to correct misspelled words. The text is then tokenized by splitting words by whitespace. After tokenization a lemmatizer is applied. This strips each token of their endings preserving just the stem of the word. Both the tokens and lemmas are then Part-of-speech (POS) tagged using the default NLTK POS-tagger, which is trained on the Penn Treebank

tagset (Marcus, Santorini & Marcinkiewicz, 1993), in order to determine the word class of each unit. The POS-tagger categorizes the tokens and lemmas into one of 36 categories (e.g. adjective or adverb). The default POS-tagger currently uses an averaged perceptron tagger (Duran et al., 2019). It is also possible to apply the Stanford POS-tagger. This is however very computationally inefficient and it has therefore been decided by the authors to rely on the default NLTK tagger. The data is then saved as both an individual data frame for each conversation and one concatenated data frame containing all conversations.

The next phase is where the alignment measures are actually calculated. This is done in two ways; on a turn-by-turn basis and on a conversation basis. In both cases lexical and syntactic alignment are calculated in a two-step process. First, the utterances are converted into bi-grams and the frequency of each bigram is calculated for each turn in the conversation. The frequencies are represented in a vector form. Secondly, from the vectors the cosine similarity is used to calculate alignment. This method takes cosine to the angle of the vector. Advantages of this method are that: 1) It normalises the length of the utterances making it possible to compare turns of different lengths to one another, and 2) the resulting outcome is on a scale from 0 to 1, where higher scores indicate more alignment. This makes the interpretation simple (Duran et al., 2019). The calculation of the syntactic alignment has one additional element to it. Repetition of lexical bigrams are removed in order to reduce the inflation that happens in syntactic alignment from just repeating the same phrase.

For the semantic alignment the lemmas are converted into a high dimensional semantic vector based on pretrained word embeddings. The pretrained word embeddings are created from the Google News Corpus using Gensim's implementation of the word2vec algorithm (Rehurek & Sojka, 2010) and consists of 3 million 300-dimensional word vectors. When calculating a vector for a target word the neighbouring words are treated as positive examples, whereas randomly sampled words from a lexicon is treated as negative examples. A neural network is then trained on distinguishing between these two cases and the weights of the network (in this case 300) are used to create a vector for each word (Jurafsky & Martin, 2019). To get the overall semantic value of each utterance in our data, the vector for each lemma in the utterance are added together. The semantic alignment can then be estimated by calculating the cosine similarity between a vector of the current and previous utterance (Duran et al., 2019).

Alignment has been found to correlate with task structure (Fusaroli et al., 2013). However, the current study was interested in the alignment which is not due to task structure and we therefore subtracted the 'contextual alignment' with the alignment generated by the participants. There are multiple ways to do this. Two commonly used methods are to create either a shuffled baseline or surrogate pairs. In the shuffled baseline the turn order is shuffled within participants. This removes the temporal effects of alignment maintaining only alignment which can be ascribed to the context (e.g. the use of directional terms in a social maze task). Surrogate pairs involve mixing interlocutors together, that did in fact not interact with one another. In this baseline the turn order is preserved. Hence, this baseline captures alignment that happens over time due to for instance particular words being used across dyads due to task structure. In general, the surrogate baseline is thought to be more conservative and is therefore used in the ALIGN package and in this project (Duran et al., 2019). Having a conservative baseline is especially important due to the different characteristics of the three datasets.

Analysis of alignment

Different analyses were carried out on the turn-by-turn dataset in order to answer the three hypotheses presented. We used the syntactic and lexical scores calculated on the lemmas for all analyses to not distinguish between different forms of the same word. All models were run in brms (Bürkner, 2018) using 2 chains with 2000 iterations (1000 warm-up). Ideally the models would be run with more chains and iterations but due to limitations on computational capacity this was not done. This of course leaves some uncertainty regarding the estimates.

<u>H1: Baseline analysis</u>

First, we created three baseline beta regression models, one for each dataset, combining all alignment measures in order to determine whether alignment in each condition was different from the baseline (surrogate pairs). As the lexical and syntactic alignment were not normally distributed a zero one inflated beta distribution was used. We also included the conversations as a random intercept and turn-order (time) as a random slope in order to account for shared variance in turns of the same conversation. A normal(0,1) prior was applied to the betas in order to help the model converge. The general structure of the models were as follows:

$$Syn, lex, sem \sim Intercept + baseline + (1 + scale(time)|conversation),$$
$$family = ZeroOneInflatedBeta$$

H2 + H3: Interaction type differences in Human alignment

The three datasets were then combined into a single data frame containing both the alignment scores of real and surrogate pairs. Using a beta regression model, alignment was modelled as a function of the type of interaction, turn-order (time) and an interaction between these two. Again, the conversations were modelled as a random intercept and turn-order (time) as a random slope. Only alignment from human participants were included in this part of the analysis to investigate how different types of interlocutors affect human tendencies to align. To simplify the model we modelled the human-human condition as intercept. The surrogate pairs were included in the final model in an attempt to account for differences in baseline alignment across datasets. The structure of the model was as follows[1]:

$$Syn, lex, sem \sim baseline + WoZ + Com + scale(time) + scale(time) * (WoZ + Com) + (1 + scale(time)|conversation), family = ZeroOneInflatedBeta$$

Results (Line)

Hypothesis 1)

The baseline analysis clearly indicated that there was an increase in alignment from surrogate to real pairs as none of the beta estimate distributions overlapped with zero. The size of the difference did however vary both based on alignment type (e.g. the difference for semantic alignment are larger than the difference for syntactic alignment) and across datasets. The beta estimates can be found in Table 2.

***Table 2**: Beta estimates of real pairs compared to surrogate pairs from baseline model*

Interaction Type	Lexical (95 % CI)	Syntactic (95 % CI)	Semantic (95 % CI)
Human-Human	0.30 (0.25:0:34)	0.10 (0.07:0.13)	0.22 (0.20:0.24)
Human-WoZ	0.33 (0.25:0.40)	0.07 (0.03:0.10)	0.17 (0.13:0.20)
Human-Computer	0.49 (0.07:0.97)	0.11 (0.00:0.21)	0.49 (0.39:0.60)

Hypothesis 2):

The results suggested differences in alignment dependent on the type of interaction. All model estimates and 95% Bayesian Credibility Intervals (CI) can be found in table 1 in appendix. For semantic alignment, the results showed a negative effect of the HWI condition

1 WoZ = Human alignment to the Wizard of Oz, Com = Human alignment to the computer

Simon A. M. Hansen - 201609506
Line Kruse - 201608877

with high certainty compared to the HHI intercept (the 95% Bayesian CI did not overlap with zero). Further, the model estimated a positive effect of the condition HCI compared to the HHI intercept, however, with slightly more uncertainty. For syntactic alignment results showed a positive effect of both the HWI and HCI condition compared to the HHI intercept, with a small uncertainty in the latter case. Lastly, the HWI condition had a negative effect on lexical alignment compared to the HHI intercept, with a small uncertainty, while the HCI condition had a positive effect with high certainty. Thus, results suggested that semantic alignment was greatest in HCI and lowest in HWI, syntactic alignment was greatest in HHI and lowest in HWI, and lexical alignment was greatest in HCI and lowest in HWI. Posterior distributions for all three alignment measures for each type of interaction are plotted in figure 1.

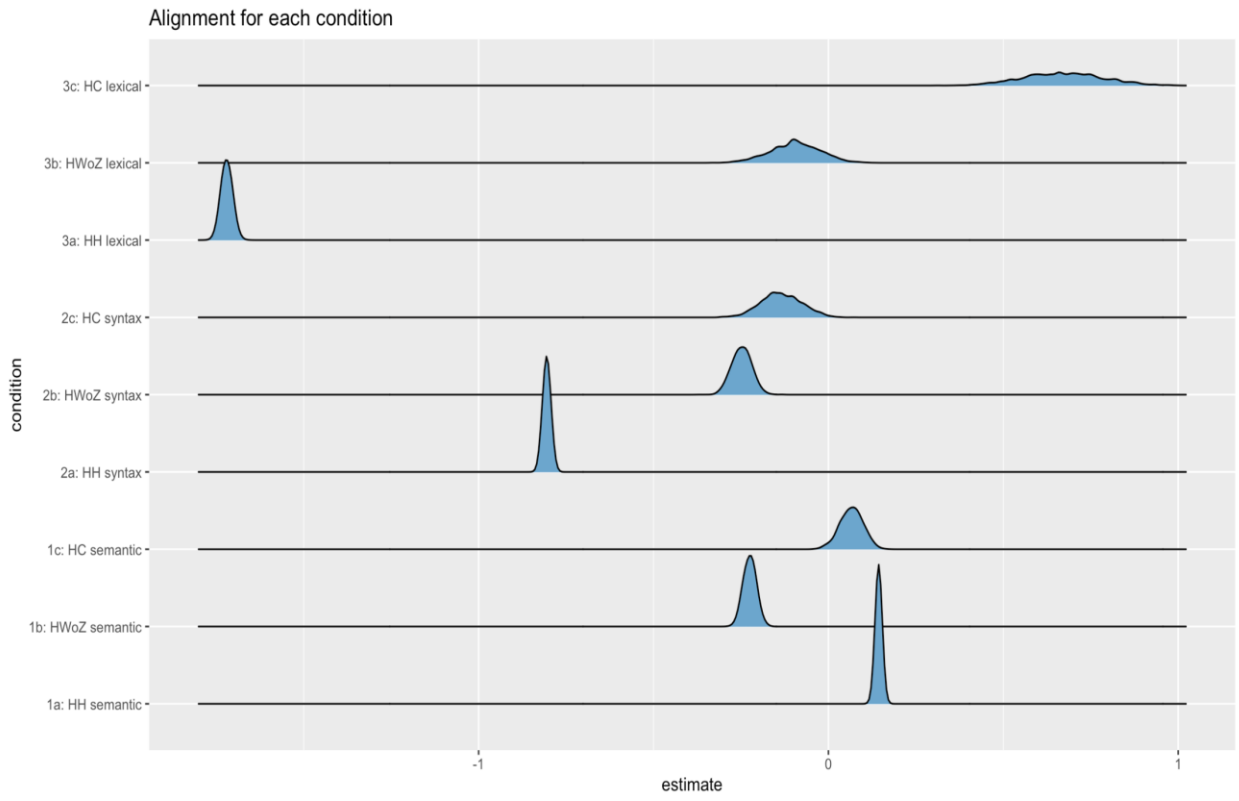

***Figure 1**: Estimate for each type of alignment for each type of interaction. HHI condition is the intercept, and the remaining posterior distributions visualize the difference in alignment compared to the intercept. An estimate of zero indicate no effect of the condition, estimates above zero indicate a positive effect, and estimates below zero indicate a negative effect.*

Simon A. M. Hansen - 201609506
Line Kruse - 201608877

Hypothesis 3):

The model showed an interaction effect between time and type of interaction (HHI, HWI, HCI), indicating that the effect of time differed between the three conditions. All model estimates and 95% Bayesian CI can be found in table 2 in appendix. For Human-Human interactions the effect of time was estimated to be negative for semantic and syntactic alignment, and positive for lexical alignment. All estimates had high certainty (the 95% Bayesian CI is under zero). For Human-WoZ interactions the effect of time was estimated to be negative for all three alignment measures. However, the 95% Bayesian CI overlapped slightly with zero for semantic and syntactic alignment, while lexical alignment had high certainty. For Human-Computer interactions the effect of time was estimated to be positive for all three alignment measures. In this case, only semantic alignment had high certainty, while the posteriors for syntactic and lexical alignment overlapped slightly with zero. In sum, the model was fairly confident that semantic and syntactic alignment decrease over time while lexical alignment increases over time for the HHI condition, all three alignment measures decrease over time for the HWI condition, and all three alignment measures increase over time for the HCI condition. The linear fits of alignment predicted by time for each type of interaction are plotted in figure 2.

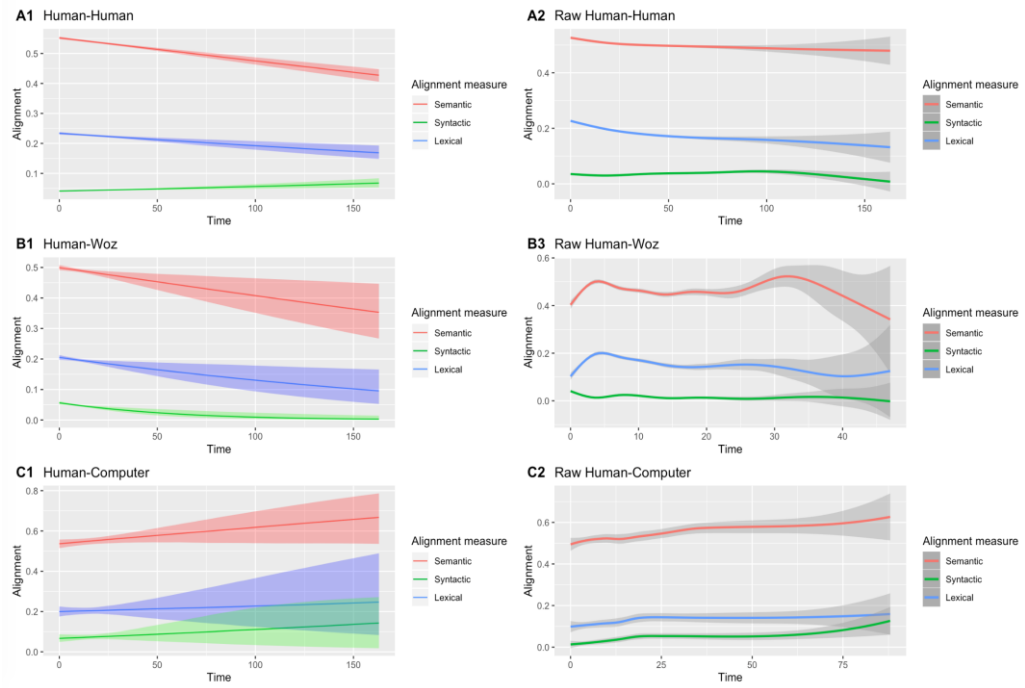

***Figure 2**: Alignment as an effect of time. Left: model fit. Right: Raw data. A: Human-Human interactions. B: Human-WoZ interactions. C: Human-Computer interactions. Red: semantic alignment. Green: Syntactic alignment. Blue: lexical alignment.*

Discussion (Line)

The analysis showed that each alignment measure was different from the baseline (surrogate pairs), in line with H1. Thus, the level of alignment was greater than what could be ascribed to chance, task structure or other confounds given to the experimental setup. Further, the results indicate that semantic and lexical alignment was greatest when participants interacted with a computer, while syntactic alignment was greatest when two humans were interacting with each other. For all three measures, alignment was weakest when participants interacted with a Wizard of Oz. H2 is hence only partly supported. Lastly, alignment appeared to decrease over time in Human-Human and Human-WoZ interactions, only with the exception of lexical alignment, which increased over time in HHI. Contrary, alignment increased over time in Human-Computer interactions, supporting H3.

The observation that the amount of alignment differed between linguistic levels across the three types of interactions, may suggest that each level of alignment could be differentially sensitive to context. Although the particular task structure was accounted for, the context in which participants interacted still differed in whether it was spoken or written dialogue, as well as the distance between partners (phone or computer). People appeared to align more semantically and lexically, when interacting with a computer compared to another human, which may indicate that these are more susceptible to strategic control than syntactic alignment. Syntactic alignment may be too difficult to employ strategically to compensate for disruptions in the conversation, and could compose a more automatic process compared to the other two. The finding that alignment was lowest in HWI, indicate that the belief about the partner's capabilities may not have influenced participant's linguistic behaviour to the extent previously suggested. If that was the case, we would have expected the level of alignment to be similar to the HCI. Rather, as the WoZ had equal linguistic capabilities as a human partner, the results suggest that alignment was primarily driven by the actual capabilities of the partner.

This interpretation is supported by the third analysis. The fact that human interlocutors align more over time when interacting with a computer is in line with previous findings, and could suggest that the linguistic behaviour of humans is influenced by their prior belief about the computer's capabilities, which is reflected in a tendency to reuse more of the expressions used by the computer to increase the likelihood of understanding. However, the results could

alternatively reflect that the interaction was actually less successful, and that people attempted to compensate for problems of misinterpretation. The notion of a synergy indeed implies a reciprocal compensation mechanism, in which interlocutors readily react to perturbations in order to preserve function (Fusaroli et al., 2013). Hence, interaction with a computer may simply induce more perturbations to the "system". Even top chatbots make more mistakes than humans, as they may fail to comprehend and produce an appropriate answer to the input they receive. The fact that alignment in the WoZ condition develops more comparably to the human condition indicates that the increase in alignment in HCI is more likely to be due to such compensation to perturbations rather than prior beliefs about the system's capabilities. If this proposition is true it would be possible to use alignment measures as a measure of chatbot success. If the partner aligns more than the computer we might assume that the chatbot is not successful in its communication. Interestingly, the analyses indicate that the greatest and most informative difference between HCI and HHI may not be in the amount of alignment as much as in the development of alignment over time.

N-grams (Simon)

The current analysis of both lexical and syntactic alignment was based on bigrams, however, the ALIGN package further allows the use of unigrams, trigrams etc. The choice of n-gram applied has implications for the resulting alignment score. In general, moving from unigrams to n-grams increases our ability to model complex structures and dependencies between words. However, using a high *n* carries the risk of overfitting to the data. In the current analysis we would imagine that lexical and syntactic alignment scores would decrease as *n* increases, as it becomes harder to detect similar n-grams. While the current study relied on the default of the ALIGN package, one might attempt to find the size of n-grams that best capture relevant information in the data and use that as the unit of analysis. We could even imagine combining different n-gram models to capture different levels of complexity in alignment.

Turn-by-turn analysis (Simon)

The ALIGN package measures alignment on a turn-by-turn basis meaning that it calculates the cosine similarity between two vectors representing one turn and its previous turn. We might suspect that alignment occurs on a slower time scale as well (e.g. a person might reuse a word or sentence structure that their interlocutor used 10 turns ago). This is not captured by

the analysis and further investigation could try to create a radius that incorporates x number of turns and calculates the mean cosine similarity to all vectors in the given radius.

Limitations (Simon)

While the datasets share some similarities (e.g. no physical interaction, dialogue between 2 partners) there are also multiple differences between them. Four problems have been identified: 1) the number of conversations vary, 2) the length of conversations vary, 3) the 'task' varies across dataset, and 4) the modality of communication varies across dataset. All these factors are potential confounds that might affect the level of alignment. It is also generally acknowledged that spoken and written language are quite different (e.g. Akinnaso, 1982; Redeker, 1984) and therefore it seems fair to assume that this may affect alignment to some extent. By including the baseline level of alignment (surrogate pairs) into the analysis we tried to model away some of these dataset differences but it should be clear that we are probably only succeeding in doing so to some extent. Further studies should aim to compare more compatible datasets.

Ethical considerations (Simon)

In general, few ethical concerns are associated with the presented study. Two of the three datasets have been constructed for research purposes meaning that the participant were knowingly participating in the study and were asked to give consent to the use of the data. The last dataset was collected during the Loebner Prize 2019 competition and we assume that the 'judges' were informed that the transcript of their conversations would be uploaded. In addition, all data were anonymised. Another important ethical consideration is whether our application/study has a dual-purpose element. Given that the current study consists of an alignment analysis it is hard to see how this can be applied to do harm.

Further studies (Simon)

The current studies used text as data, which limits what types of alignment you are able to assess. Using audio files would have allowed us to investigate alignment in measures such as prosody, speech rate and articulation. Given that such alignment measures are relevant in applications such as Siri, which are speech based, it makes sense to try to study these alignment behaviours in human-human interaction in order to allow for more natural communication with digital assistants.

References

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and speech*, *25*(2), 97-125.

Bell, L., Gustafson, J., & Heldner, M. (2003, August). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS* (Vol. 3, pp. 833-836). Citeseer.

Brennan, S. E. (1991). Conversation with and through computers. *User modeling and user-adapted interaction*, *1*(1), 67-86.

Branigan, H., & Pearson, J. (2006). Alignment in human-computer interaction. *How people talk to computers, robots, and other artificial communication partners*, 140-156.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41-57.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., Nass, C. I., & Hu, J. (2004). Beliefs about mental states in lexical and syntactic alignment: Evidence from human–computer dialogs. *Proceedings of the CUNY Conference on Human Sentence Processing*. College Park, MD: University of Maryland

Bürkner P (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395–411. doi: 10.32614/RJ-2018-017

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychological methods*. *24*(4), 419–438

Godfrey, J. & Holliman, E. (1993) Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/LDC97S62

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2013). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147-157.

Jurafsky, D. & Martin, J. H. (2019). Chapter 6: Vector semantics and Embeddings. In *Speech and Language Processing* (3rd ed. draft). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Koulouri, T., Lauria, S., & Macredie, R. D. (2016). Do (and say) as I say: linguistic adaptation in human–computer dialogs. *Human–Computer Interaction*, *31*(1), 59-95.

Maes, A., Marcelis, P., & Verheyen, F. (2007). Referential collaboration with computers. *Anaphors in text: cognitive, formal and applied approaches to anaphoric reference*, *86*, 49.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.

Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., & Nass, C. I. (2006, April). Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1177-1180). ACM.

Pickering, M. J., & Garrod, S. (2014). Interactive alignment and language use. *The Oxford handbook of language and social psychology*, 131-140.

Radlinski, F., Balog, K., Byrne, B., & Krishnamoorthi, K. (2019). Coached conversational preference elicitation: A case study in understanding movie preferences. Retrieved from https://research.google/tools/datasets/coached-conversational-preference-elicitation/

Redeker, G. (1984). On differences between spoken and written language. *Discourse processes*, *7*(1), 43-55.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Schmader, C., & Horton, W. S. (2019). Conceptual Effects of Audience Design in Human–Computer and Human–Human Dialogue. *Discourse Processes*, *56*(2), 170-190.

Wang, Y., Reitter, D., & Yen, J. (2014, June). Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics* (pp. 55-62).

Worswick, S. (2019). *Mitsuku - Loebner Prize 2019*. Retrieved from http://www.square-bear.co.uk/mitsuku/loebner2019.htm

Appendix

***Table 1**: Estimate and 95% Bayesian CI for each type of alignment for each type of interaction. H-H condition is the intercept, and the remaining estimates indicate the difference in alignment compared to the intercept. An estimate of zero indicate no effect of the condition, estimates above zero indicate a positive effect, and estimates below zero indicate a negative effect.*

Semantic	Estimate	CI lower	CI upper
Intercept (HH)	0.14	0.13	0.16
WoZ	-0.22	-0.26	-0.19
Computer	0.07	0.00	0.13
Syntax			
Intercept (HH)	-0.81	-0.82	-0.79
WoZ	-0.25	-0.30	-0.19
Computer	-0.13	-0.24	-0.02
Lexical			
Intercept (HH)	-1.72	-1.75	-1.69
WoZ	-0.09	-0.24	0.04
Computer	0.67	0.46	0.89

***Table 2**: Estimate and 95 % Bayesian CI of the effect of time for each type of alignment. H-H condition is the intercept, and the remaining estimates indicate the difference in alignment compared to the intercept. An estimate of zero indicate no effect of the condition, estimates above zero indicate a positive effect, and estimates below zero indicate a negative effect.*

Semantic	Estimate	CI lower	CI upper
Time	-0.05	-0.06	-0.04
Time:WoZ	-0.01	-0.06	0.03
Time:Com	0.11	0.04	0.18
Syntax			
Time	-0.05	-0.07	-0.03
Time:WoZ	-0.06	-0.13	0.02
Time:Com	0.08	-0.08	0.24
Lexical			
Time	0.07	0.03	0.10
Time:WoZ	-0.41	-0.61	-0.21
Time:Computer	0.06	-0.26	0.40