Conversational entrainment in the use of discourse markers

Štefan Beňuš

Constantine the Philosopher University, Nitra, Slovakia & Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

Abstract. Entrainment is the tendency for participants in conversations to develop behaviour similar to one another in multiple dimensions. The degree of such entrainment is linked to the emotional state and empathy of the speakers and people who entrain to their conversational partners are seen as more socially attractive, likeable, competent, more intimate, and the interactions with such partners as more successful. It is thus important that ICT interfaces for supporting wellbeing and empathy employ also some module of entrainment.

In this paper we analyze entrainment in the acoustic, prosodic and pragmatic domains connected to the use of Slovak discourse marker 'no' in the spoken modality of task-oriented collaborative dialogues. We analyze how speaking behaviour changes due to interacting with a different partner, and consequently, how entrainment is employed. We use acoustic and prosodic information extracted from the signal and labelled pragmatic functions of the marker (including acknowledgment, backchannel, reservation, topic shift, etc.). Results suggest a varied picture with both entrainment and disentrainment present in the data. Regarding the relationship between entrainment in acoustic-prosodic features and more cognitively complex features of pragmatic meaning and discourse functions, we found both matches and mismatches between the two.

Keywords: Entrainment, prosodic features, human-computer interaction, well-being, discourse markers

1 Introduction

Speech entrainment is the tendency of interlocutors to become similar to each other in terms of their acoustic and prosodic production and relates to cognitive and social aspects of communication and information transfer. Some aspects of speech entrainment appear to be almost automatic, take place early in the interaction, and presumably thus employ lower levels of the cognitive communication systems (e.g. [7, 14, 15]). Other aspects might require higher cognitive functions since they include linguistic encoding (e.g. [17] for a review). Moreover, entrainment observable in spoken modality may be linked in non-trivial ways to entrainment in gestures, body postures, and other aspects of visual modality.

adfa, p. 1, 2011.

Social aspects of spoken entrainment include findings that humans perceive conversational partners who entrain to their speaking style as more socially attractive and likeable, more competent and intimate, and conversations with such partners as more successful (see [10] for a review of extensive literature). It has also been shown that humans may consciously decrease their similarity to others in order to increase their social distance to the interlocutor or to show a negative attitude toward the interlocutor. It is thus important that ICT interfaces for supporting wellbeing and empathy employ also some module of entrainment.

Importantly for social robotics, not only do humans entrain to other humans, but studies have shown that they also entrain to computer systems, and that subjects do adapt to machines similarly to human conversational partners (e.g. [1, 6, 11, 19]). A better understanding of entrainment is thus important for all applications in human-machine communication that rely on Spoken Dialogue Systems. Due to the naturalness of the spoken modality for humans, human-robot interactions are likely to rely heavily on speech and the social aspects of these interactions will play a major role in the advances in the field of social robotics. Hence, the ability to mimic the tendency for entrainment in human-human conversation is important for human-robot conversation as well, if social robotics systems are to be as natural and effective as human partners.

In order to contribute to these future advances in implementation and engineering of human-computer interactions, we analyze here entrainment patterns in the usage of a single discourse marker in human-human task-oriented dialogues. Primarily, we are interested in comparing patterns in the single marker to general entrainment. Knowledge gained form such a comparison is potentially usable in at least two ways. First, observed similarities might allow faster processing, since entrainment can be assessed on a small sample of data. Second, observed differences might point to differences in cognitive processes underlying these two types of entrainment. Furthermore, we compare entrainment in acoustic-prosodic features with a limited investigation of entrainment in cognitively more demanding functional characteristics of this discourse marker. Section 2 describes all methodological aspects of data collection, labelling, extraction, and analysis. Section 3 presents our results that are discussed and summarized in Section 4.

2 Methodology

2.1 Corpus

Data for this study were selected from a corpus of dialogues in which two interlocutors were playing collaborative games, i.e. tasks in which spoken interaction is required to achieve successful completion. These games were adapted from the OBJECT Games described in [8, 9] and the current corpus is also described in [2]. Briefly, interlocutors could not see each other and were seated in a quiet room facing a computer screen and using a mouse. One player - the describer – verbally depicted the position of a target image in relation to other images on her screen. The second player – the placer – was supposed to place the same image to a position as close as possible to the position of the image on the describer's screen. Players were encouraged, and motivated with a small reward, to match the positions perfectly and their success was measured a 100-point scale based on how closely the pixel-positions of the two objects matched. Each dialogue consisted of 14 tasks and the roles of the describer and the placer were switched repeatedly and at the end were equally divided between the two players.

The corpus used for this analysis consists of six dialogues in Slovak involving seven native speakers (3 females 4 males). Importantly, five players (LP, KM, DF, MD, VR) participated in the recording twice (with a different partner) and two male subjects played only one game. The corpus contains almost four hours of speech (3h, 54m), there are 21773 words in total, and 2371 unique words.

2.2 Discourse marker 'no' in Slovak

It has been widely observed that discourse structuring of interactions is signalled and facilitated by the distribution and prosodic characteristics of discourse markers. They not only display the discourse structure but play a prominent role in creating it; see e.g. [18] for a review. We concentrate on discourse marker '*no*' in Slovak [2]. This form might represent a shortening of affirmative particle *áno*, which means 'yes' in Slovak. It can thus typically signal many functions of *okay* identified and analyzed in [8] such as backchannel, acknowledgment, beginning of a new discourse segment, or agreement. Additionally, '*no*' can signal non-commitment and mild disagreement since Slovak *no* is also a conjunction roughly meaning 'but'.

2.3 Features: labelling and extraction

In this paper we concentrate on three types of features: acoustic-prosodic characteristics contained within the discourse marker 'no' itself, acoustic-prosodic features characterizing the speech of an interlocutor as a whole, and pragmatic features capturing the discourse communicative functions of 'no'. The unit of analysis will be the individual task; recall there are 14 tasks for each dialogue.

Acoustic and prosodic features in the signal were extracted using Praat [3]. We extracted a standard set of acoustic features including duration and the slope of F0 (only for no-tokens), and mean, maximum, minimum of F0 and Intensity, and also voice quality features such as jitter, shimmer, harmonics-to-noise ratio or spectral tilt for both no-tokens and each task of the game. Since in this paper we compare the behaviour of the same speaker in her two sessions, features were not normalized.

Table 1 shows the scheme for labelling the discourse functions of 'no'. The scheme was designed following the scheme used for the functions of 'okay' in [9] and appended for additional unique features of Slovak '*no*': Z, J, or RZ [2].

Table 1. Scheme for labelling discourse functions of 'no', adapted from [9]

Label	Meaning	Label	Meaning

R	I acknowledge that I understand, I got it	**H**	Hesitation, I am stalling for time
RP	I acknowledge that I understand, and please continue	**E**	I want to repair/redo something I’ve just said or did
RZ	I acknowledge that I understand, but I want to add something or express mild disagreement	**PH**	I express an assessment of something that has just happened, usually after receiving a score
RN	I acknowledge that I understand, and I want to start a new topic or a new discourse segment	**J**	I Soften of what is to follow, a hedge
N	I want to start a new topic or a new discourse segment	**K**	I signal the end of the current topic or discourse segment
S	I agree, also as an answer to a questions, usually meaning *yes*	**D**	I encourage some action, go on, do something
Z	I want to express an idea opposite to the one implied before, usually meaning ‘but’ or ‘well’	**?**	None of the labels correspond to the perceived meaning

2.4 Analyzing entrainment

We analyze here the differences in the speaking behaviour of the five subjects as a function of their conversational partner. In other words, we ask if a subject changed his/her speaking behaviour between the two games that s/he played, and if the answer is positive, we ask if his/her behaviour was more similar or dissimilar to that of his/her conversational partner. To answer this question for discrete dependent variables, such frequency of ‘*no*’, we run a chi-square test comparing the frequencies of the speaker in her two games. To assess entrainment for continuous features, we run a t-test for the difference in the feature values extracted from the two games of the speaker. For both discrete and continuous features, if a significant difference is reported, we calculate the difference of the means of feature F in the two games played by the speaker, and the difference between the means of F extracted from the two interlocutors that played the game with the target speaker. If the signs of the two values are identical, we take it as evidence of entrainment. As an illustration, a speaker has mean pitch of 200Hz in session$_1$ and 230Hz in session$_2$. Her interlocutor in session$_1$ has mean pitch 180Hz and the other partner in session$_2$ 220Hz. Both differences (200-230 and 180-220) have the same sign (irrespective of the order of the operands), which corresponds to the target speaker adjusting her mean pitch to be more similar to her interlocutor. If the signs are different, we refer to this as dis-entrainment: the speaker changed her behaviour but became less similar to her interlocutor.

As an additional measure of entrainment in the use of ‘*no*’ we follow [13] and compute entrainment as the negative value of the absolute difference between the frequency of ’*no*’ words between the two interlocutors (S1 and S2) shown in (1) below; *count* corresponds to the number of no-words and *ALL* to the sum of all words other than ‘*no*’.

$$ENTR(no) = -\left|\frac{count_{S1}(no)}{ALL_{S1}} - \frac{count_{S2}(no)}{ALL_{S2}}\right| \quad (1)$$

Hence, the lower (more negative) the ENTR(no) value, the less entrainment is there between the two interlocutors.

3 Results

We start with analyzing entrainment in the frequency of using '*no*' employing ENTR(no) measure in (1) above for the five speakers who played the game twice with a different partner. Table 2 shows the values for this measure separately for each of the five speakers and session. Columns 3 and 4 show that the highest entrainment was reached in the games played by speaker DF; identical numbers here correspond to game partners (DF, for instance, played her games with KM and VR). All other games are characterized by a rather low entrainment in this feature.

Table 2. Entrainment in the frequency of 'no' usage

Speak er	**Gender**	**ENTR(no) with partner**		**ENTR(no) with self**	**X^2**	**p**
LP	M	-2.6	-2.08	-0.83	2.9	0.1
KM	F	-0.61	-3.01	-1.6	4.7	0.03
DF	F	-0.61	-0.55	-0.85	2.7	0.1
MD	M	-2.05	-3.01	-0.32	0.7	0.4
VR	F	-2.08	-0.55	-2.32	16.2	0.001

The fifth column of Table 2 reports the values of our entrainment measure from the data for a single speaker in the two games s/he played. We assume that if ENTR(no) is low, the speaker differed in his/her frequency of '*no*' usage between the two games, and thus potentially s/he entrained or dis-entrained to his/her conversational partner. The last two columns show the results of chi-square tests assessing the significance of the difference in the no-usage for a speaker in his/her two games. We see that two speakers (KM and VR) significantly changed their no-usage and for two speakers (LP and DF) a tendency was reported. Further examination following the rationale described in section 2.4 revealed that both former speakers (KM and VR) significantly dis-entrained from their partners while the two latter speakers (LP and DF) showed a tendency for entrainment.

Finally, we checked if there is a significant difference between ENTR(no) in the third and fourth columns on the one hand and the fifth column on the other. In other words, we tested if speakers entrained more to themselves or to their partners, and we expected the former prediction will be born out. We do have a small number of values (5 and 10 respectively), and the difference with this data is not significant; however, the direction of the effect follows the expectation (greater entrainment for self, t[11] =

-1.1, p = 0.15). Doubling the data yields an almost significant value (t[24] = -1.63, p = 0.058). This result provides a sanity check and suggests that assessing entrainment with this measure is plausible.

The discourse function of ‘*no*’ signalling that the speaker acknowledges and understands the previous utterance and wishes for his/her partner to continue (RP in Table 1) was the most frequent at 31% of all no-tokens. We examined if speakers (dis)entrain not only on the frequency of no-usage but also on this pragmatic function of the marker. For this purpose we calculated the frequency of RP function from all the uses of ‘no’ per speaker and session and followed the same steps and chi-square tests as described above. Our results show that three speakers (DF, KM, and VR) differed significantly in their frequency of RP function among no-uses. Following the method described above, the first two speakers disentrained and the last one entrained. Hence, interestingly, a speaker (VR) might disentrain in no-frequency but entrain in the frequency of a particular discourse function; we also have a speaker with the opposite pattern (DF). This supports the idea that entrainment on more cognitively complex features might differ from other types of entrainment.

We follow with testing (dis)entrainment in terms of acoustic and prosodic features extracted from no-tokens. Table 3 shows how many of our five speakers either entrained to their partner in terms of a given acoustic feature (2nd column), disentrained from the partner (3rd column), or did not produce a significant difference between the two games played (4th column) in terms of features extracted from no-tokens.

Table 3. Number of speakers showing entrainment, disentrainment, or no difference in the two games played

	No-tokens			**Entire corpus**		
Feature	**Speakers differ in 2 games**		**Speakers do not differ in 2 games**	**Speakers differ in 2 games**		**Speakers do not differ in 2 games**
	Entr	**Disentr**		**Entr**	**Disentr**	
Intensity mean	3	1	1	4	0	1
F0_mean	0	2	3	0	3	2
F0_slope	0	0	5			
Duration	1	1	3			
F1	1	1	3			
F2	0	1	4			
jitter	0	1	4	2	0	3
shimmer	2	0	3	3	0	2
hnr	1	1	3	2	1	2
spec. tilt	1	2	2	0	3	2
Total	**9**	**10**	**31**	**12**	**10**	**18**

Table 3 shows that the most pervasive pattern is no significant change between the behaviour of a speaker in the two games s/he played. Significant changes are equally distributed between more similarity and dissimilarity to the conversational partner.

It might be the case that the situation in terms of entrainment on the acoustic-prosodic features of no-tokens either reflects a general pattern of entrainment in the corpus, or that the patterns of entrainment in no-tokens and in a conversation as a whole are different. To approach this issue, we examined entrainment on the features in Table 3, but this time considered all data from the corpus and took a single task (14 tasks in a game) as a unit of analysis. Hence, for most cases, we had 14 data points per speaker and game. The results are in the three rightmost columns of Table 3. We observe that the results from the overall entrainment assessment are very similar to the ones reported for no-tokens only. Non-significant differences between the two games played by a speaker are most pervasive with roughly equal distribution of entrainment and disentrainment. In terms of features, speakers tend to be more inclined to entrain on intensity and voice quality and less on other features. This is to be expected since entrainment is most likely to occur for features that are most redundant and thus carry a minimal functional load in terms of linguistic contrasts; e.g. [16].

Finally, expecting different behaviour based on the task role (Describer vs. Placer), as reported for other features in corpora of task-oriented games (e.g. inter-speaker intervals in map-tasks [4]), we also tested entrainment for the two roles separately. The results for both no-tokens and entire corpus suggest that being in a more dominant role (Describer) induces 1) greater tendency for entrainment and 2) greater tendency to differ in the two games.

4 Discussion and conclusion

Our small-scale pilot study revealed two main findings. First, we observed less entrainment than expected. Studies of entrainment in similar corpora suggest pervasive entrainment (e.g. [11]) while in our corpus, entrainment is present but not overwhelmingly. This might be due to a different language/culture (Slovak vs. English), less sophisticated measures of entrainment in this study, or several other differences. Second, the situation in just no-tokens is very similar to the entire corpus. This opens the possibility for more efficient ways of accessing well-being of users from the degree of entrainment to the system from a smaller set of target tokens.

Our results also show that the crucial feature of models of entrainment suitable for implementation in automatic systems interacting with humans is their adaptability. This is because speakers employ their own individual strategies for entrainment in prosodic and voice quality features, some speakers or features do not participate in entrainment, and some show disentrainment. Moreover, differences were reported for the same speaker between entrainment on the basic word-usage and its discourse function. Hence, entrainment is highly subject-dependent, which is an important finding for applications in human-computer interaction since accommodation to the user needs to be personalized. In future we plan to employ more sophisticated measures of entrainment, different units of analysis, and more speakers. For example, we will compare adjacent inter-pausal units (IPUs) across a turn-exchange and compare with random pairs of IPUs from the two speakers.

Acknowledgements. This research was supported by a VEGA grant 2/0202/11 and results also from of the project implementation: RPKOM, ITMS 26240220064 supported by the Research & Development Operational Programme funded by the ERDF.

5 References

1. Bell, L., Gustafson, J., Heldner, M.: Prosodic adaptation in human-computer interaction. *Proceedings of 15th ICPhS,* 2003.
2. Beňuš, Š: Prosodic forms and pragmatic meanings: the case of the discourse marker '*no*' in Slovak, *Proceedings of CogInfoCom*, 2012.
3. Boersma, P., Weenink, D.: Praat: doing phonetics with computer. [www.praat.org]
4. Bull, M., Aylett, M.: An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. *Proceedings of ICSLP*, 1998.
5. Chartrand, T., Bargh, J.: The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology* 76:893-910, 1999.
6. Coulston, R., Oviatt, S., Darves, C.. Amplitude convergence in children's conversational speech with animated personas. *Proceedings of ICSLP*, 2002.
7. Delvaux, V. & Soquet, A.: The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*, 64, 145-173, 2007.
8. Gravano, A., Hirschberg, J., Beňuš, Š.: Affirmative cue words in task-oriented dialogue, *Computational Linguistics*, 38(1), 1-39, 2012.
9. Gravano, A., Benus, S., Chávez, H., Hirschberg, J., Wilcox, L.: On the role of context and prosody in the interpretation of *okay*. *Proceedings of ACL*, 800-807, 2007.
10. Hirschberg, J.: Speaking More Like You: Entrainment in Conversational Speech. *Proceedings of Interspeech*, 27-31, 2011.
11. Levitan, R., Hirschberg, J.: Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of Interspeech*, 2011.
12. Nass, C. Moon, Y., Fogg, B. J., Reeves, B, Dryer, D. C.: Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43(2), 223-239, 1995.
13. Nenkova, A., Gravano, A., Hirschberg, J.: High frequency word entrainment in spoken dialogue. *Proceedings of ACL/HLT*, 169-172, 2008.
14. Nielsen, K.: Implicit Phonetic Imitation is constrained by Phonemic Contrast. *Proceedings of the 16th ICPhS*, 2007.
15. Pardo, J.: On phonetic convergence during conversational interaction. Journal of Acoustical Society of America, 119(4), 2382-2393, 2006.
16. Pentland, A.: To Signal Is Human. *American Scientist*, 98, 204-210, 2010.
17. Pickering, M., Garrod, S.: Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169-226, 2004.
18. Redeker, G.: Review article: Linguistic markers of linguistic structure. *Linguistics*, 29(6), 1139–1172, 1991
19. Stoyanchev, S., Stent, A.: Lexical and syntactic priming and their impact in deployed spoken dialogue systems. *Proceedings of NAACL*, 2009.