Annotation specifications of a dialogue corpus for modelling phonetic convergence in technical systems

Grażyna Demenko, Jolanta Bachan

Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland
lin@amu.edu.pl, jolabachan@gmail.com

Abstract: The present paper describes spoken dialogue corpus creation and its annotation specification for analysis and objective evaluation of phonetic convergence in human-human communication. The analysis of the corpus will serve for creation of convergence models which could be implemented in spoken dialogue systems based on spontaneous, expressive speech. The corpus consists of 13 hours of dialogues between 16 pairs of Polish native speakers and controlled dialogues with a teacher. The speakers knew each other and were at similar age, but during the recording could not see each other. In each recording session the pair of speakers conducted 4 dialogues in neutral scenarios and 6 dialogues in expressive scenarios, 3 dialogues with the teacher, 2 repetition tasks and 1 reading, which provided about 1 hour of speech for each pair. The corpus is being annotated on several layers: orthographic transcription of text, prosody, noise, flow of speaking turns, dialogue acts, agreement and disagreement intervals, extraordinary events and speaker's attitude. This scenarios combination and annotation specifications are novel, and promise to provide an empirical foundation for both linguistic and computational dialogue modelling of both face-to-face and man-machine dialogue. The results of preliminary analyses were used for selection of recording scenarios for German speakers. The next step of the ongoing project is to record dialogues between Polish L1 speakers with German L1 and Polish L2 speakers.

1 Introduction

Phonetic convergence in a dialogue is a natural phenomenon. The notion of phonetic convergence is related to the Communication Accommodation Theory (CAT) which regards interpersonal conversation as a dynamic adaptive exchange that was established in the 1970s [7, 8]. Phonetic convergence in dialogue involves adaptation of segmental and suprasegmental features of speech to those of the interlocutor, with the function of cooperatively or manipulatively signalling social common ground [10]. The main assumption of this theory is that interpersonal conversation is a dynamic adaptive exchange involving both linguistic and nonverbal behaviour between two human interlocutors. The phenomenon of inter-speaker accommodation in spoken dialogues is well-known in psycholinguistics, communication and cognitive sciences [6]. The features that undergo accommodation include lexical, syntactic, prosodic, gestural and postural features, as well as turn-taking behaviour [11]. The function of inter-speaker accommodation is to support predictability, intelligibility and efficiency of communication, to achieve solidarity with, or dissociation from, a partner and to control social impressions. The significant role of such adaptive behaviour in spoken dialogues in human-to-human communication has important implications for human-computer interaction. In the context of speech technology applications, communication accommodation is important for a variety of reasons: models of convergence can be used to improve the naturalness of synthesised speech (e.g. in the context of spoken dialogue systems, SDS), accounting for accommodation can improve the prediction of user expectations and user satisfaction/frustration in real time (in on-line monitoring) and is essential in establishing a more sophisticated interaction management strategy in SDS applications to improve the efficiency of human-machine interaction.

Communicative adaptation has been viewed as a potential functionality in human-machine interaction, but the phenomenon of phonetic convergence has not yet been exploited in human-machine communication systems in detail (cf. [2, 9, 12]), for which an appropriate corpus is needed. The present paper presents creation of a corpus of spoken dialogues with a special focus on annotation specifications created for analysis of phonetic convergence between the interlocutors. In the future the corpus will be used to create quantitative models of accommodation phenomena exhibited in specific properties of speech (acoustic-prosodic, temporal and spectral) in human and human-computer dialogues in view of their implementation in speech technology.

2 Corpus design

The corpus is being created within an ongoing project which aims at (1) extracting phonetic features which can be mapped on to a synthetic signal, (2) creating dialogue models applicable in human-machine interaction and (3) practical evaluation of the types and degree of phonetic convergence. It is planned to record dialogues with different configuration of speakers' L1 / L2:

- Polish L1 speaker with Polish L1 speaker
- Polish L1 speaker with German L1 / Polish L2 speaker
- German L1 speaker with German L1 speaker
- German L1 speaker with Polish L1 / German L2 speaker

The present paper describes only the creation of the dialogue corpus between the Polish L1 speakers. The recordings of the dialogues between the other groups is planned as the next step of phonetic convergence analysis, also in different language pairs.

2.1 Subjects

For the corpus, 16 pairs of speakers were recorded: 8 male-male pairs and 8 female-female pairs who knew each other and/or were close friends. From all the subjects such metadata was collected as: name, sex, age, height, weight, education, profession, information on languages spoken and proficiency levels.

The youngest subject was 19 years old and the oldest was 58 years old (recorded in pair with a 50-year-old), the biggest age difference was 12 years and the average age difference was 3 years. Only 3 pairs of female speakers were above 30 years old, all the other subjects were younger than 29 years. The average age of the subjects was 27 years. Additionally, in each session a 33-year-old female teacher carried out 3 dialogues with each of the subjects.

2.2 Scenarios

2.2.1 Controlled scenarios

There were 3 tasks in the *controlled* scenarios. In the first task, the subject heard a recording of a short sentence over the headphones by a male of a female speaker and the subject's task was to repeat the sentence in a way to best imitate the melody of the original. The sentence "Jola lubi lody" (Eng. "Jola likes ice-creams") was played 6 times with a stress on different syllables: "**Jo**la lubi lody" or "Jola **lu**bi lody" or "Jola lubi **lo**dy".

The second task was to read a dialogue. In the third task the subject was to read/repeat the phrases of the same dialogue, but imitating the melody of phrases of the pre-recorded speech (a similar task as in the first one, but this time the sentences were longer and their expressiveness differed).

These controlled recordings were carried out to evaluate general speakers possibilities to produce segmental and suprasegmental structures (accent type and placement, consonant cluster production) and to assess whether the speakers had talent to imitate other's speech and whether they could be expected to phonetically converge with the other speaker to a great extent. For these scenarios only orthographic and prosodic annotation was applied at this stage, which made it impossible to compare the accentuation and voice similarity between the speakers and the teacher (e.g. using Dynamic Time Warping (DTW) method [5]). The annotation on the phone level will be performed in the next stage of annotation.

While recording the corpus, two phoneticians carrying out the recordings assessed perceptually that one speaker had little tendency to adjust his speech to the speech recordings.

2.2.2 Neutral scenarios

The neutral scenarios consisted of 4 dialogues. The first was a decision-making dialogue in which the interlocutors were to decide together what to take to a desert island to *survive*. They could choose 5 items from a given list. This was a cooperative dialogue, there was to be no role asymmetry and the maximum convergence was expected.

The second dialogue was based on a *diapix* task [14] where in a cooperative dialogue the subjects were to find 3 differences between two pictures. There was no role asymetry and the subjects had to describe their pictures in order to find the differences between the subjects' pictures.

The last two dialogues from the neutral scenarios were *map-tasks*. One of the speakers was asked to play a tourist in a foreign city and the other was to pretend to be a receptionist in a hotel. The tourist was calling the hotel at which he booked a room to ask how to get there. The subjects had the map of the city to be used in the dialogue. There was asymetry in the dialogue and it was expected that the tourist would converge to the receptionist, i.e. the leader of the dialogue. The map-task was recorded twice with the speakers exchanging their roles.

2.2.3 Expressive scenarios

The set of expressive dialogues was divided into 4 groups: a) asymetry: power – dominant vs. submissive (entertainment scenario), b) asymetry: emotionally coloured speech – valence: positive vs. negative (fun vs. sadness/fear, terrorist attact scenario), c) no role asymetry: both speakers in agremeent vs. both speakers in disegrement (provocation in art) and d) dialogues with the teacher (also agreement and disagreement).

In the first scenario one of the speakers played a role of a tourist information centre assistant of a big city and his task was to provide information about events and interesting places in the city and to convince the caller to choose at least his offer. If he had convinced the caller, the assistant would have received an *award* from his boss. The other person was a party-goer who wanted to to find out what attractions the city offered at night. The dialogue was asymmetric, designed to boost a strong convergence to the tourist information assistant, the leader of the dialogue, who showed great enthusiasm. The same scenario was used again, but with the exchanged speakers roles.

In the following scenario, the tourist information assistant was informed about *terrorist attacks* in the city and was unwilling to provide any information about the entertaining events in the city. Despite the threat of another attack, the assistant has to inform the caller about the interesting places in the city, but the best procedure was to suggest only the safest options or to convince the caller to stay at home. The other speaker was again the party-goer who despite the threat of terrorist attacks wanted to go out to have some fun. The dialogue was to show a strong asymmetry and convergence to the assistant, the leader, who showed no enthusiasm to provide any information and even scared the caller that going out might put his life in danger. After the dialogue was finished, the subjects changed their roles and carried out a similar dialogue again.

Dialogues on provocation in art were designed to elicit mutual convergence as there was to be no role asymetry. The subjects saw pictures of very provocative content and their tasks were to discuss them and *approve* this form of art in the first scenario, and later they both were asked *oppose* and condemn such art. The same set of approve/oppose dialogues was also carried out between each subject and the teacher.

Finally, the last dialogue between the teacher and the subject was about Madonna's provocative performance. Both parties strongly supported their own views. The teacher – the *opponent* – was very conservative and thought Madonna was evil and condemned Madonna for crucifying herself during her concert. Contrary, the subject – the *supporter* – was a fan of modern art, liked provocations and loved Madonna. Their task was to exchange their opinions of the presented photo from Madonna's concert.

The dialogues wiht the teacher allowed to control the dialogues, boost more expressiveness if needed, more fun or extreme indignation. The teacher could also control the length of the dialgoues and make it longer if she thought the given subject did not speak long enough.

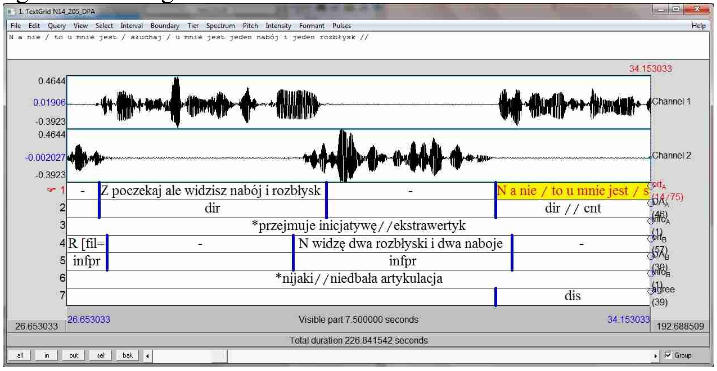
2.3 Recording session

The recording session started by signing the agreement by the subjects to take part in the recordings and let their voice be used for research purposes. Each session lasted approximately 2 hours and altogether 13 hours of recordings were collected. The recordings were carried out in a professional studio. The speakers could hear each other over headphones, but could not see each other. For the recordings, 2 overhead microphones and 2 stationary microphones were used, providing 4 mono channels of recordings, 2 for each speaker, at 44.1 kHz sampling frequency. The software used for the recordings was Cakewalk Sonar X1 LE and Roland Studio Capture hardware was the audio interface employed.

3 Annotation specifications

The annotation of the dialogues was carried out on 7 layers in Praat [3]:

1. ort_A – orthographic and prosodic annotation, speaker A
2. DA_A – dialogue acts, speaker A
3. info_A – metadata: information about speaker, e.g. excited, information about relation between speakers, e.g. dominant, any additional information, speaker A
4. ort_B – orthographic and prosodic annotation, speaker B
5. DA_B – dialogue acts, speaker B
6. info_B – metadata, speaker B
7. agree – parts of dialogues where both speakers agree or not, information about convergence in dialogue.


1. TextGrid N14_Z05_DPA
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
34.153033
0.4644
0.01906
-0.3923
Channel 1
0.4644
-0.002027
-0.3923
Channel 2
1 - Z poczekaj ale widzisz nabój i rozbłysk - N a nie / to u mnie jest / s
2 dir dir / / cnt
3 *przejmuje inicjatywę / / ekstrawertyk
4 R [fil= - N widzę dwa rozbłyski i dwa naboje -
5 infpr infpr
6 *nijaki / / niedbała artykulacja
7 dis
ort_A (14/75)
DA_A (46)
info_A (1)
ort_B (57)
DA_B (39)
info_B (1)
agree (39)
26.653033
26.653033
Visible part 7.500000 seconds
34.153033
192.688509
Total duration 226.841542 seconds
all in out sel bak
☑ Group

N a nie / to u mnie jest / słuchaj / u mnie jest jeden nabój i jeden rozbłysk //

Figure 1 – Sample dialogue annotation on 7 layers in Praat.

The annotation layers are described in details in the following sections. The annotation process is still in progress and so far only a few recordings sessions were fully annotated according to the presented specifications. The sample annotation is presented in Figure 1.

3.1 Orthographic segmental and prosodic suprasegmental annotation

The orthographic and prosodic annotation tier is the richest tier. The dialogue is divided into speaker's turns and transcribed orthographically. The white space is the word boundary and change of speakers, a longer pause (or filled pause) between two stretches of speech indicates the necessity of inserting a time boundary on the annotation tier. Numbers (times, dates, etc.) are spelled in their spoken form (e.g. "thirty four", "fifth of January"). Abbreviations and acronyms are represented by their full forms of spelling. The punctuation rules do no apply – the commas or full stop are replaced by the prosodic markers. The text is written in lower letters, apart from proper names of people, institutions, city names, etc. In the text, the speaker noises are marked as fillers (label: [fil]) or breaths (label: [spk=b]) or laughter ([laugh]) or any other noise coming from the speaker ([spk], e.g. click, cough). Words coming from other languages than Polish are transcribed with a language label specifying the origin of the word, e.g. [len=EN] welcome. If the language is not recognised, then the label [len] is entered with two asterisks ** that follows. The mispronounced words are marked with an asterisk (e.g. _*armchair_ if it is pronounced as /ˌɑː(ɹ)mˈʃeə(ɹ)/) and words or stretches of speech that are completely unintelligible are transcribed by a sequence of two asterisks: **. Words with prolonged pronunciation (e.g. speaker thinks what to say) are annotated with the mark = at the beginning of a word. There are two types of noises coming from the outer sources which are annotated, these are: stationary noise (label [sta]) and intrusive noise ([int]). Stationary noise is a background noise of a a stable loudness amplitude over some time and intrusive noises are short and loud noises which typically occur only once (like a door slam). The parts of recordings which cannot be used for the analysis are marked with the label [trash].

3.1.1 Suprasegmental prosodic annotation

In the description of prosodic phenomena, the following factors are taken into consideration: prominence of syllables and two levels of prosodic phrase boundary strength - the weak and strong boundaries. In addition, elements of discourse with high impact on the prosodic structure of speech are taken into account [5].

Table 1 - Summary of specification of perceptual annotation of prosody.

Label	Usage	Placement	To what refers
+	strong syllable prominence	after emphasised vowel	a single word
/	weak phrase boundary	after the last word in the phrase	a single word or a group of words
//	strong phrase boundary - falling intonation, declarative sentence		
//?	strong phrase boundary - rising intonation, interrogative sentence		
//!	strong phrase boundary - falling intonation, exclamation sentence		
$	grammatical indication of phrase boundary but prosodic cues are not unequivocal		
{	ungrammatical boundary (speaker inserts a pause/breath/filler within the utterance)	after the word sequence where the ungrammatical boundary occurs	
/..	apposition	after the last word in a phrase	
/@	backchannel		
/~	incomplete utterance cut off at the end		
~/	incomplete utterance cut off at the beginning	before the first word in a phase	
!	clear emphasised word	after the word with emphasis	a single word

The annotation of prominence and phrase boundaries are guided by both meaning, i.e. the syntactic, semantic and discourse cues, and the acoustic features of speech. In order to reconcile the two criteria, (1) labels marking weak phrase boundaries were introduced (boundary type /) in places where syntactically and semantically such a boundary occurs, but the acoustic cues are very subtle, and (2) labels indicating ungrammatical phrase boundaries (type \) which are clearly marked by prosody, but appear in "unexpected" locations from the point of view of the semantic, syntactic and/or discourse structure of an utterance. Boundaries marked as $ are placed where syntactically and semantically boundaries are likely to occur, but the speaker does not realize them at the acoustic level. The strong phrase boundaries (boundary type //) corresponds to final stops in punctuation rules, and depending on the sentence type, the boundary may be followed by a question mark (type //?) or an exclamation (type //!). The proposed specification of prosodic annotation is summarised in Table 1.

3.1.2 Flow of speaking turns

The annotation of flow of speaking turns is based on Allen's interval algebra [1]. The speaking turn relations are based on their occurrence in time as well as their meaning. For example, if a new topic appears in a dialogue after a longer pause, then the analysis of the flow is "reset" and the new turn may be marked as "A before B". The flow is marked at the beginning of a turn before the orthographic transcription of the text, with one of the labels describing Allen's relations. The set of relations are presented in Table 2.

Table 2 – Dialogue flow annotation based on Allen's interval algebra [1].

N	Relation	Descripion	Label		N	Relation	Descripion	Label
1	_____ ... A	speaker A starts dialogue	SA		12	____ ____ A B	B meets A	S
2	_____ ... B	speaker B starts dialogue	SB		13	______ A ____ B	A overlaps with B	N
3	____|____ A A	two utterances divided by a time boundary	*none*		14	______ ____ A B	B overlaps with A	N
4	____ ____ A A	two utterances divided by a pause	*none*		15	_____ __A______ B	A starts with B	R
5	____|____ B B	two utterances divided by a time boundary	*none*		16	_________ ____A B	B starts with A	R
6	____ ____ B B	two utterances divided by a pause	*none*		17	____ ____A_____ B	A during B	T
7	____ A ____ B	A before B	P		18	_________ _A_ B	B during A	T
8	____ ____ A B	B before A	P		19	____ _______A_ B	A finishes B	J
9	____ A ____ B	B after A, semantically the turn relates to A's last utterance	Z – graphically the same as 7		20	_________ A____ B	B finiehes A	J
10	____ ____ A B	A after B, semantically the turn relates to B's last utterance	Z – graphically the same as 8		21	_________ ____A____ B	A is equal to B, B is equal to A	E
11	____ A ____ B	A meets B	S					

3.2 Dialogue acts

For dialogue act annotation, 20 dialogue acts were selected from DIT++ Taxonomy [4]. For each dialogue act, an acronym was created to speed up the annotation process. The selected list of dialogue acts includes: allo-feedback, auto-feedback, commissives, contact management, directives, information providing (confirm and disconfirm), information seeking, open meeting, own communication control, partner communication management, social obligations management functions (salutation, self-introduction, apologising, gratitude, valediction), time management, topic shift announcement and turn management (like turn take, turn accept, turn grab). Each utterance could be classified with more than one dialogue act function.

3.3 Metadata and additional information

The metadata and additional information layer about the speakers consisted of:

- speaker's personality assessment – speaker's attitude annotated with an asterisk * before the word (extrovert, introvert, dominant, subordinate, shy, funny, joker, nervous, insecure, neuter, neutral, ...)
- information about relation between speakers – information whether the speakers cooperate, whether they want to reach the common goal or whether one of the speaker's ignores the other
- extraordinary events (repetitions, stuttering, cursing, hyper-correct pronunciation, ...).

3.4 Agreement/Disagreement parts of dialogue

The agreement/disagreement layer of dialogue divided the dialogue into intervals in which the speakers were in accordance or were quarrelling. The annotation was made on perceptual analysis and was based on the meaning of the dialogue – only when the annotators were sure the speakers did not agree on some topic, then the disagreement interval was inserted.

4 General assumptions for annotations in scenarios

The general assumptions for annotations in scenarios were as follows:

1. Controlled scenarios – only orthographic segmental and suprasegmental prosodic annotations, annotation on phone level planned in the future for which automatic segmentator for Polish SALIAN [13] will be used.
2. Neutral scenarios – orthographic segmental and suprasegmental annotations, metadata and additional information (speaker's personality assessment, evaluation of relations between speakers, extraordinary events), agreements/disagreement parts of dialogue.
3. Expressive scenarios – rich annotation on each level like in 2: neutral scenarios.

5 Discussion and future work

This scenario combination and annotation specifications are novel, and promise to provide an empirical foundation for both linguistic and computational dialogue modelling of both face-to-face and man-machine dialogue.

Testing the developed annotation based on phonetic-acoustic analyses will provide basis for its possible application in technical systems like speech synthesis and automatic recognition of spontaneous speech. Additionally, we expect to find specific individual segmental and suprasegmental features which potentially could be useful for speaker characterization.

6 Acknowledgements

The present study was supported by the Polish National Science Centre, project no.: 2014/14/M/HS2/00631, *"Automatic analysis of phonetic convergence in speech technology systems"* and was conducted in cooperation with the project partner Prof. Bernd Möbius who is the leader of a project *"Phonetic convergence in Human-Machine Communication."*

References

[1] Allen, J. F.: Maintaining knowledge about temporal intervals. In: *Communications of the ACM*. ACM Press. pp. 832–843, ISSN 0001-0782, 26 November 1983.

[2] Bachan, J.: *Communicative alignment of synthetic speech*. Ph.D. Thesis. Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland, 2011.

[3] Boersma, P. and D. Weenink: PRAAT, a system for doing phonetics by computer. In: *Glot International* 5(9/10), pp. 341-345, 2001.

[4] Bunt, H: Dialogue pragmatics and context specification. In: Bunt, H. and W. Black (Eds.): Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics, pp. 81 – 150, Amsterdam: John Benjamins 2000.

[5] Demenko, G.: *Korpusowe badania języka mówionego.* (Polish: Corpus studies of spoken language). Akademicka Oficyna Wydawnicza EXIT, 2015.

[6] Dogil, G.: *Language talent and brain activity* (Vol. 1). Walter de Gruyter, 2009.

[7] Giles, H.: *Accent mobility: A model and some data*. Anthropological Linguistics 15, pp. 87 – 105, 1973.

[8] Giles, H., N. Coupland and J. Coupland: Accommodation Theory: Communication, context, and consequence. In: Giles, H., J. Coupland, and N. Coupland (Eds.): *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pp. 1 – 68, Cambridge University Press 1991.

[9] Lelong, A. and G. Bailly: Study of the phenomenon of phonetic convergence thanks to speech dominoes. In: A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt. *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*, Springer Verlag, pp. 280-293, 2011, LNCS AI. <hal-00603164>

[10] Pardo, J. S.: On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119, pp. 2382 – 2393, 2006.

[11] Pickering, M. J. and S. Garrod: Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, pp. 169 – 225, 2004.

[12] Savino M., Lapertosa L. Caffò A., and M. Refice: Measuring prosodic entrainment in Italian collaborative game-based dialogues. *Proceedings of the 18th International Conference on Speech & Computer* (SPECOM 2016), Budapest 23-27 August 2016, p.476-483, LNCS Series n.9811, pp.476 – 483, Springer Verlag, 2016.

[13] Szymański, M. and S. Grocholewski: Transcription-based automatic segmentation of speech. In: *Proceedings of 2nd Language and Technology Conference*, Poznań, pp. 11 – 14, 2005.

[14] van Engen, K. J., M. Baese-Berk, R. E. Baker, A. Choi, M. Kim and A. R. Bradlow: *The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profile*s. Language and Speech, Vol. 53, 4, pp. 510 – 540, December 2010.