






The Harmonia Corpus – A Dialogue Corpus for Automatic Analysis of Phonetic Convergence

Jolanta Bachan¹ (✉) , Mariusz Owsiany² , and Grażyna Demenko¹ 

¹ Faculty of Modern Languages and Literatures, Adam Mickiewicz University
in Poznań, al. Niepodległości 4, 61-874 Poznań, Poland
{jbachan, lin}@amu.edu.pl

² Poznan Supercomputing and Networking Center, ul. Jana Pawła II 10, 61-139 Poznań, Poland
mowskianny@man.poznan.pl

Abstract. The work presents the creation of a dialogue corpus for analysis and formal evaluation of phonetic convergence in spoken dialogues in human-human and human-machine communication, with the goal of comparing dialogue features at all levels of language use. The Harmonia corpus was created within a project which aims at (1) extracting phonetic features which can be mapped on a synthetic signal, (2) creating dialogue models applicable in a human-machine interaction and (3) practical evaluation of the convergence. For the corpus the following language groups were recorded: 16 pairs of Polish speakers speaking Polish (native speech), 10 pairs of German speakers speaking German (native speech), 12 pairs of German and Polish speakers speaking Polish (non-native speech), and 10 pairs of Polish and German speakers speaking German (non-native speech). The speakers could hear each other, but could not see each other. The recording scenarios consisted of controlled, neutral and expressive tasks and provided over 27 h of speech. This scenario combination is novel and promises to provide an empirical foundation for both linguistic and computational dialogue modelling of both face-to-face and man-machine dialogue.

Keywords: Dialogue corpus · Phonetic convergence · Recording scenarios · Human-computer interaction

1 Introduction

Phonetic convergence in a dialogue is a natural phenomenon. Phonetic convergence involves shifts of segmental as well as suprasegmental features in pronunciation towards those of a communicative partner [22]. The research on this phenomenon has its origin in the Communication Accommodation Theory (CAT) that has been established in the 1970s [14, 15]. 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. This theory started as a model of interpersonal communication and has since been developed to encompass insights from a number of disciplines, including linguistics, sociology and psychology. One central ingredient of CAT is the

attention that speakers and listeners direct at the speech of their interlocutors. Individual adjustments to speech are assumed to subserve the function of controlling (maintaining, reducing or increasing) social distance. The speaking style of conversational partners thus converges, diverges or remains unchanged, depending on the strategies applied by the interlocutors. Most studies in the CAT framework aim at finding social motivations for accommodation behaviour and share the assumption that the processes underlying the manipulation of speech behaviour are – at least partially – under the speaker’s conscious control.

Speakers accommodate their behaviour on semantic, lexical, syntactic, prosodic, gestural, postural and turn-taking levels [24]. 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.

Studies on phonetic convergence rest on the assumption that the incoming speech signal undergoes an early, front-end analysis, which decomposes the speech signal into a set of features. In principle, each feature can be the target of convergence processes in production. Acoustic features investigated include (e.g. [3, 9, 12, 16, 31]): voice-onset time (VOT), formants, voicing, F0 range and register, pitch accents, intensity, duration, pausing, and speaking rate, as well as the long-term average spectrum (LTAS). Such acoustic measures can be complemented by perceptual judgements of the presence or degree of convergence.

Communicative adaptation has been viewed as a potential functionality in human-machine interaction to improve system performance [1, 6, 10, 11, 26, 27]. It can be assumed that a responsive human-computer interface that accommodates some features of the human interlocutor may be perceived as more user-friendly and may even lead to enhanced learning. The phenomenon of phonetic convergence that occurs naturally and partly automatically in human-human communication has not yet been exploited sufficiently in human-machine communication systems and the manipulation of the phonetic structure of speech generated in SDS environment with the aim of converging to the human speech pattern has been hardly investigated so far (cf. [2, 4, 18, 19, 21, 28]).

Apart from being an information exchange, it is widely recognised that human conversation also is a social activity that is inherently reinforcing. As such, new conversational interfaces are considered social interfaces, and when we participate in them we

respond to the computer linguistically and behaviourally as a social partner. Human-computer interfaces that mimic human communication (and thus account for accommodation/convergence phenomena) will constitute next-generation conversational interfaces for speech technology applications. The benefits of using speech as an interface are multiple: simplicity (speech is the basic means of communication), quickness, robustness, pleasantness (related to social aspects of spoken communication, building relations), convenience (it can be used in hands-free and eyes-free situations or when other interfaces are inconvenient), it can be used as an alternative interface for the disabled and has some technical benefits (readily available hardware such as a telephone is sufficient).

Although the literature on communication accommodation in spoken dialogues in human interaction is fairly extensive, research on human-computer interaction has yet to face the challenge of investigating whether users of a conversational interface likewise adapt their speech systematically to converge with a computer software interlocutor. At this moment, the application of phonetic convergence in speech technology applications is not feasible for two reasons. The first one is related to the lack of an efficient quantitative description of this complex behavioural phenomenon as it occurs in spoken language. Past research on interpersonal accommodation has focused on qualitative descriptions of the social dynamics and context involved in linguistic accommodation. It also has relied on global correlation measures to demonstrate linguistic accommodation between two interlocutors. Only quantitative predictive models that account for the magnitude and rate of adaptation of different features, the factors that drive dynamic adaptation and re-adaptation, and other key issues will be valuable in guiding the design of future conversational interfaces and their adaptive processing capabilities. The second reason is that current SDS architectures are not designed to accommodate natural dialogue with human users, therefore a platform for testing quantitative models of inter-speaker accommodation does not yet exist.

The present paper describes creation of the Harmonia spoken dialogue corpus for analysis and objective evaluation of phonetic convergence in human-human communication. In Sect. 2 the corpus design is presented: the information about the subjects, the reading and repetition tasks, the scenarios of the dialogues, and the recording setup in a professional studio. Section 3 outlines the annotation specifications of the corpus. The last section concludes the paper and presents works carried out on the Harmonia corpus.

2 Corpus Design

The dialogue corpus Harmonia was created within a project which aimed at (1) extracting phonetic features which could be mapped on a synthetic signal, (2) creating dialogue models applicable in human-machine interaction and (3) practical evaluation of the types and degree of phonetic convergence. The Harmonia dialogue corpus contains dialogues with different configuration of speakers' L1/L2:

- subcorpus Harmonia_PL1_PL1: Polish L1 speaker with Polish L1 speaker
- subcorpus Harmonia_PL1_DE1_PL2: Polish L1 speaker with German L1/Polish L2 speaker
- subcorpus Harmonia_DE1_DE1: German L1 speaker with German L1 speaker

- subcorpus Harmonia_DE1_PL1_DE2: German L1 speaker with Polish L1/German L2 speaker

The subcorpus of dialogues between Poles is the biggest and the richest, containing a wider range of dialogue scenarios and complex annotation. The subcorpora of German dialogues of native and non-native speech, and the Polish dialogues between a German and a Pole are much smaller and the annotation is simpler.

2.1 Subjects

Polish L1 Speaker with Polish L1 Speaker

For the native Polish subcorpus, 16 pairs of Polish 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 the following metadata was collected: 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/phonetician carried out 3 dialogues with each of the subjects as a confederate.

Polish L1 Speaker with German L1/Polish L2 Speaker

For the non-native Polish subcorpus, 12 pairs of native Polish speakers with native German speaking Polish were recorded: 6 male-male pairs and 6 female-female pairs. All the speakers spoke Polish fluently, but their command differed a lot: one speaker had lived in Poland only for 5 months, 4 speakers lived in Poland for 19–30 years, some speakers were born in Germany in Polish families. Because of the General Data Protection Regulation (GDPR) no additional information about the subjects was collected. The confederate in the non-native Polish subcorpus was the same teacher as in the native Polish dialogues.

German L1 Speaker with German L1 Speaker¹

For the native German subcorpus, 10 pairs of German speakers were recorded: 3 male-male pairs and 6 female-female pairs. The metadata collected from the subjects included information about: age, gender, height, weight, mother tongue (for all subjects it was German), highest school-leaving qualification or university degree, job or field of study, languages and language level, region and city of childhood/youth, a note if the speakers knew each other or were strangers and information if they knew the confederate or not, and additional annotation about the course of the recording.

The youngest participant was 19 years old and the oldest was 55. The biggest age differences were 21 and 36 for two pairs, the other pairs belonged to the same age group

¹ The German dialogues between the German native speakers and Polish L1/German L2 speakers were recorded at Saarland University by the Phonetic group led by Prof. Dr. Bernd Möbius who was the partner of the Harmonia project.

around 22 years old. Each subject talked also to a confederate – a 21-year-old German student trained to carry out the task.

German L1 Speaker with Polish L1/German L2 Speaker

For the non-native German subcorpus, 10 pairs of native German speakers were recorded with Poles speaking German: 1 male-male pair and 9 female-female pairs. The metadata collected about the subjects was the same as for the native German group. The youngest participant was 19 years old and the oldest was 57. It was hard to find Polish students speaking German in Germany where the recordings took place, so the age of subjects varied a lot. The mean age of the participants was 27 years old, while the mean age difference between the speakers was 12 years. All the subjects were also recorded in a dialogue carried out with a confederate – the same German student as in the native German subcorpus.

2.2 Scenarios for Pairs: Native Speakers of Polish Speaking Polish

The recording session was composed of a few short tasks. The first tasks were controlled reading and repetition. These tasks were introduced to assess the speakers's talent to adapt their speech to the model voice and their expressiveness while reading an enthusiastic interview with a music star. The next set of dialogues were task-oriented (neutral): either the dialogues were cooperative with no leader or in the dialogue the leader was specified and it was expected that the interlocutor would adapt to the leader's voice. Additionally, a set of expressive scenarios was recorded. These dialogues were also cooperative with no leader in the dialogue when recorded in pairs of common speaker, but when each of the speakers was to talk with the teacher/phonetician, it was expected that the speaker would adapt to the teacher in their expressiveness, liveliness and language.

Such a choice of scenarios was made to apply the developed convergence models to speech technology scenarios at different kinds of call centers, automatic information services or computer games.

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 or 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: "Jola **lubi** lody" or "**Jola** lubi lody" or "Jola lubi **lody**". Figure 1 shows the short sentence uttered by the male or female speaker, with the stress marked by "+" on different syllables. The blue line on the spectrogram is the fundamental frequency (F0) of the speakers – the lower for the male and the higher of the female.

The second task was to read a dialogue. It was an interview by a reporter and a singer. The dialogue was constructed in such a way to contain neutral and expressive phrases with exclamations.

In the third task the subject was asked to read/repeat the phrases of the same dialogue as in the previous task, but imitating the melody of phrases of the pre-recorded speech

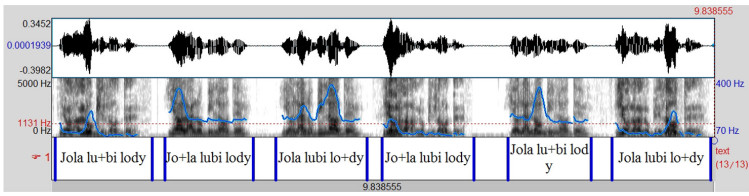


Fig. 1. The sentence “Jola lubi lody” with stress marked by “+” on different syllables.

(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. 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.

Task-Oriented Scenarios

The task-oriented (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 the following list: TV set, binoculars, matches, nails, soap, favourite teddy bear, mattress, knife, petrol, tent, pen, bowl, book, hammer, kite. 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 [30] where in a cooperative dialogue the subjects were to find 3 differences between two pictures. There was no role asymmetry and the subjects had to describe their pictures in order to find the differences. The diapixes are presented in Fig. 2. There are 10 differences between the pictures, but preliminary recordings revealed that finding all differences was taking too long and the task was simplified to finding only 3 differences.



Fig. 2. Diapixes for neutral scenario: describe and find 3 differences [30].

The last two dialogues from the task-oriented scenarios were map-tasks. One of the speakers was asked to play a tourist in a foreign city who just arrived at the main station 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 from the main station. The subjects had the map of the city to be used in the dialogue (Fig. 3). There was asymmetry 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 and with different maps.

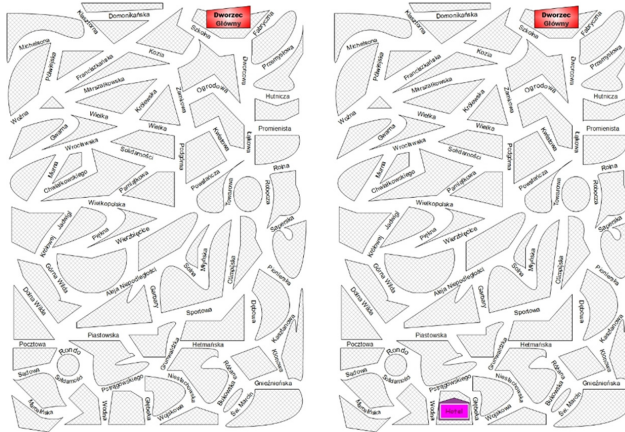


Fig. 3. Maps for the map-task: tourist's map on the left, receptionist's map on the right; "Dworzec Główny" means "Main Station."

Expressive Scenarios

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

In the first scenario one of the speakers played the 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 one of his offers. 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 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 second 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 had 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 asymmetry. The subjects saw pictures of a very provocative content and their tasks were to discuss them and approve this form of art in the first scenario, in the following dialogue they both were asked to 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, on the contrary, the subject – the supporter – was a fan of modern art, liked provocations and loved Madonna. The task was to exchange their opinions of the presented photo from Madonna's concert (Fig. 4).

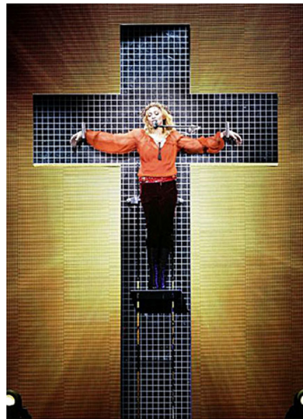


Fig. 4. Picture for the expressive scenario: Madonna on the cross [23].

The dialogues with the teacher allowed to control the course of the dialogue, boost more expressiveness in subjects if needed, add more fun or show extreme indignation. The teacher could also control the length of the dialogues and make it longer if she thought the given subject did not speak long enough in the previous tasks.

2.3 Scenarios for Pairs: Native Speaker of Polish with Native Speaker of German Speaking Polish

Controlled Scenario: Reading

The first task was for the German to read 4 sentences presented below. In the sentences were there typical for Polish fricatives, affricates, nasal sounds and consonant clusters.

Kasia zanosí koszyk z kaszą do kasy.
 Na wczasach często czytuję kiczowate czasopisma.
 Zaczął oglądać film, ale zaraz zasnął.
 Potrzeba matką wynalazku.

Expressive Scenarios

There were 2 dialogues recorded between the German and the Pole. In the first dialogue, the speakers were to agree on the presented picture (see Fig. 5). In the second dialogue, the speakers were to oppose the show by Madonna who crucified herself on the cross during her concert (see Fig. 4).



Fig. 5. Provocative art [17].

The last dialogue was recorded between the German and the Polish teacher/confederate. The teacher conducted the dialogue according to the following scenario:

- While entering a “virtual” exhibition, the subject had to repeat the 4 selected Polish sentences after the teacher (the sentences from the first task).

- Room 1: the subject talked with the teacher about provocative art, picture 1 (Fig. 5). After ca. 3-min discussion, the persons “changed the room” and the subject had to say the 4 sentences after the teacher (the sentences from the first task).
- Room 2: the subject talked with the teacher about provocative art, picture 2 (Fig. 4). After ca. 3-min discussion, on leaving the exhibition, the subject had to repeat the 4 sentences after the teacher (the sentences from the first task).

Such scenario was a source of spontaneous dialogues, but also provided three times sentences repeated by the teacher. These repeated sentences, together with the read sentences from the first task, constitute a clear material for analysis of segmental (sound) alignment of a non-native Polish speaker to the teacher in the course of discourse.

2.4 Scenarios for German Dialogues

The scenarios for German dialogues, groups (1) a pair of German native speakers and (2) a German native speaker talking to a Pole speaking German were the same and were composed of a reading task, a neutral dialogue, two expressive dialogues and a dialogue with a confederate.

Controlled Scenario: Reading

The first task was to read the following sentences including sounds which do not exist in the Polish language.

Das Leben ist eben angenehm.
 In der Nacht entfachten sie ein Feuer, es war prachtvoll.
 Ich wurde beim Essen von Nina angesprochen.
 Hörst du die Schönheit der Wörter?

Neutral Scenario

The first dialogue recorded was a decision-making task in which the interlocutors were to decide together what to take to a desert island to survive. The list of items was the same as in the Polish desert island task, but translated into German. The dialogue was cooperative and there was no role asymmetry.

Expressive Scenarios

The expressive scenarios were the same as in the Polish-German pairs speaking Polish. There were 2 dialogues recorded: in the first dialogue, the speakers were to agree on the presented picture (see Fig. 5), in the second dialogue, the speakers were to oppose the show by Madonna (see Fig. 4).

The last dialogue conducted with each of the subject was led by the confederate. The confederate was the moderator and the scenario was the same as in the non-native Polish dialogues. The subjects visited a “virtual” exhibition with two pictures: Fig. 5 and then Fig. 4. To enter the exhibition, move to another room and to leave the place the subjects had to repeat the 4 sentences in German – the same sentences as in the reading task.

2.5 Recording Session

The recordings of Polish dialogues were carried out in a professional recording studio at the Faculty of Modern Languages and Literatures, Adam Mickiewicz University in Poznań, Poland. The German dialogues were recorded at the Department of Language Science and Technology, Saarland University in Saarbrücken, Germany.

Each recording session started by signing a consent in Polish or German by the subjects to the recording their voices for the academic project purposes. Speakers answered also the questions concerning basic personal information described in Sect. 2.1 Subjects.

For the dialogue recording, the studios were specially prepared according to the highest standards [13]. In Poznań participants of the experiment felt free and could hear each other over the headphones, but could not see one another. One of the recorded persons was closed in the insulated reverberation cabin while the second speaker sat in the corner of the studio which was separated by sound absorbing panels. In Germany the recording setup was a bit different. The two speakers were sitting in a big soundproof booth separated by a thin partition wall which made it impossible to see the interlocutor, but they could hear each other.

The speech prompts were on a piece of paper, but during the recording the speakers were asked to put the paper on a small table nearby. Holding a paper is a classic source of noise and for the future recordings a music stand will be used during the recording sessions.

In Poland four professional microphones were used for recordings: 2 overhead microphones (DPA 4066 omnidirectional headset microphone) and 2 stationary microphones (condenser, large diaphragm studio microphone with cardioid characteristic – Neumann TLM 103). Microphones were plugged into the high performance audio interface Roland Studio Capture USB 2.0 equipped with 12 microphone preamps. The recordings were carried out using Cakewalk Sonar X1 LE software [29]. This setup provided 4 mono channels of recordings, 2 for each speaker, at 44.1 kHz sampling frequency and 16 bit depth. Exemplary screenshot showing the process of recording a dialogue is presented in Fig. 6. First speaker's voice was recorded in sound insulation cabin (anechoic chamber): first sound track is recorded using studio stationary microphone and the third sound track was recorded with the headset microphone. Second and fourth sound tracks concern respectively studio and headset microphones used by second speaker in the acoustically separated by sound absorbing panels corner of the studio. One recording session of Polish native dialogues lasted approximately 2 h and provided about 1 h of speech. The recordings between the Poles and the Germans lasted about 30 min. During the recordings, the speakers were asked to drink mineral water to refresh their throats. Short breaks were also taken if needed.

In Germany only 2 microphones were used – one for each speaker. This gave 2 sound tracks of recordings. The voice of the other speaker was heard in the background of the main recorded speaker, but was silent enough not to cause trouble in speech analysis.

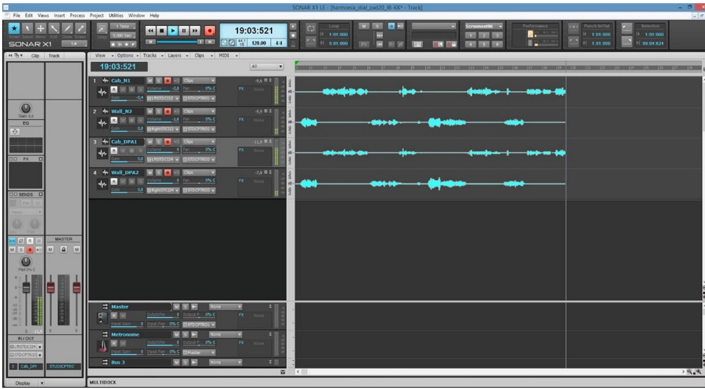


Fig. 6. Screenshot of the Sonar X1 LE recording software. 1st and 3rd sound tracks – speaker A, 2nd and 4th sound tracks – speaker B [29].

The set of scenarios for each of the subcorpus provided altogether 27.5 h of speech recordings. In the corpus 96 different speakers and two confederates were recorded. The summary of the corpus is presented below:

- Polish L1 speaker with Polish L1 speaker: 13 h, 16 pairs of speakers
- Polish L1 speaker with German L1/Polish L2 speaker: 3 h 46 min, 12 pairs of speakers
- German L1 speaker with German L1 speaker: 5 h 20 min, 10 pairs of speakers
- German L1 speaker with Polish L1/German L2 speaker: 5 h 24 min, 10 pairs of speakers

3 Annotation Specifications of the Dialogue Corpus

The first annotation specification was designed to be carried out on 7 tiers in Praat [5]:

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.

The annotation tiers were described in detail in [8] and the annotation work continued for a few weeks on the native Polish subcorpus (see Fig. 7). However, the process was very time consuming and the annotation was reduced to only two tiers 1 and 4, i.e. orthographic and prosodic annotation of speaker A and speaker B, respectively (see Fig. 8).

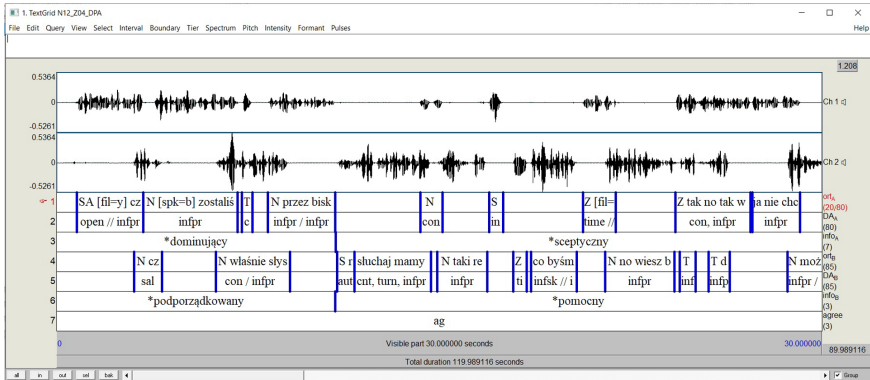


Fig. 7. Sample dialogue annotation on 7 tiers in Praat, a Polish dialogue.

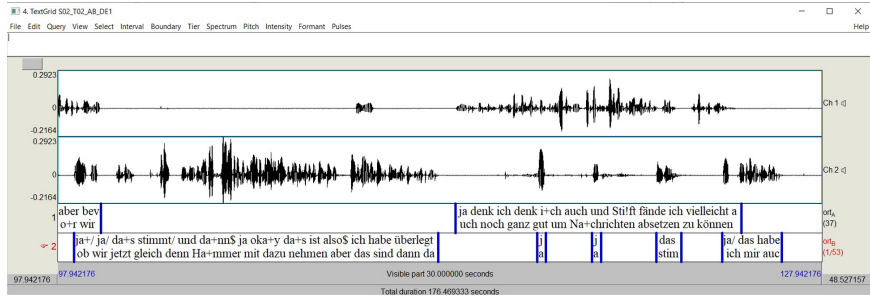


Fig. 8. Sample dialogue annotation on 2 tiers in Praat, a German dialogue.

4 Discussion and Conclusion

In the present paper, the creation of the Harmonia dialogue corpus for phonetic convergence analysis and modelling was presented. The dialogue scenarios and controlled speech prompts were shown in detail and the recording method and equipment setup in two professional studios were presented. Finally, the annotation specifications of spontaneous speech were outlined. The 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 by providing systematic quantitative data on convergence in a set of plausible scenarios.

The analysis of Polish vowels on the Harmonia corpus was presented in [20] and the preliminary analysis of segmental and lexical convergence in German dialogues between native speakers of German and native speakers of Polish was presented in [25]. Additionally, three perception tests were carried out to see if people could sense differences in speaker voices even in short fragments of recorded speech. Three factors were evaluated in those tests: pitch, tempo and meaning of speech. The tests showed that people could sense the changes in all three investigated factors. More about the analyses carried out in the Harmonia project can be found at [7]. The analysis of the corpus will

serve in the future for creation of convergence models which could be implemented in spoken dialogue systems based on spontaneous, expressive speech.

Acknowledgements. The present study was supported by the Polish National Science Centre, Harmonia 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 was the leader of a project “Phonetic convergence in Human-Machine Communication”. More about the Harmonia project can be found at: http://wczp.pl/technologie_mowy/spech_convergence.html

References

1. Bachan, J.: Modelling semantic alignment in emergency dialogue. In: Proceedings of 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, 25–27 November 2011, pp. 324–328 (2011)
2. Bachan, J.: Communicative alignment of synthetic speech. Ph.D. thesis. Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland (2011)
3. Baumann, S., Grice, M.: The intonation of accessibility. *J. Pragmat.* **38**, 1636–1657 (2006)
4. Beňuš, Š.: Social aspects of entrainment in spoken interaction. *Cogn. Comput.* **6**(4), 802–813 (2014). <https://doi.org/10.1007/s12559-014-9261-4>
5. Boersma, P., Weenink, D.: PRAAT, a system for doing phonetics by computer. *Glott Int.* **5**(9/10), 341–345 (2001)
6. Carlsson, R., Edlund, J., Heldner, M., Hjalmarsson, A., House, D., Skantze, G.: Towards human-like behaviour in spoken dialog systems. In: Proceedings of Swedish Language Technology Conference (SLTC), Gothenburg, Sweden (2006)
7. Demenko, G. (ed.): *Phonetic Convergence in Spoken Dialogues in View of Speech Technology Applications*, Akademicka Oficyna Wydawnicza EXIT, Warszawa (in press)
8. Demenko, G., Bachan, J.: Annotation specifications of a dialogue corpus for modelling phonetic convergence in technical systems. In: *Studententexte zur Sprachkommunikation - Proceedings of 28th Conference on Electronic Speech Signal Processing (ESSV)*, Saarbrücken, Germany, 15–17 March 2017 (2017)
9. Duran, D., Lewandowski, N.: Cognitive factors in speech production and perception: a socio-cognitive model of phonetic convergence. In: Matešić, M., Memišević, A. (eds.) *Language and Mind: Proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society*, pp. 15–31. Peter Lang, Berlin (2020)
10. Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. *Speech Commun.* **50**(8–9), 630–645 (2008)
11. Edlund, J., Heldner, M., Gustafson, J.: Two faces of spoken dialogue systems. In: *Inter-speech 2006*. Pittsburgh, PA, USA (2006)
12. Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., Steiner, I.: Shadowing synthesized speech – segmental analysis of phonetic convergence. In: *ISCA*, pp. 3797–3801 (2017)
13. Gibbon, D., Moore, R., Winski, R.: *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin (1997)
14. Giles, H.: Accent mobility: a model and some data. *Anthropol. Linguist.* **15**, 87–105 (1973)
15. Giles, H., Coupland, N., Coupland, J.: Accommodation theory: communication, context, and consequence. In: Giles, H., Coupland, N., Coupland, J. (eds.) *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pp. 1–68. Cambridge University Press (1991)

16. Gorisch, J., Wells, B., Brown, G.: Pitch contour matching and interactional alignment across turns: an acoustic investigation. *Lang. Speech* **55**, 57–76 (2012)
17. Hanuka, A.: Underworld. <http://www.asafhanuka.com/underground>. Accessed 02 Nov 2020
18. Jankowska, K., Kuczmariski, T., Demenko, G.: Human converging responses to natural speech and synthesized speech. *Lingua Posnaniensis* (in press)
19. Lelong, A., Bailly, G.: Study of the phenomenon of phonetic convergence thanks to speech dominoes. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. LNCS*, vol. 6800, pp. 273–286. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25775-9_26
20. Maleszewski, P.: *Analiza iloczasu polskich samogłosek w dialogach* (Eng. *Analysis of the Polish vowel length in dialogues*). MA thesis. Institute of Ethnolinguistics, Adam Mickiewicz University, Poznań, Poland (2020)
21. Oertel, C., Gustafson, J., Black, A.: On data driven parametric backchannel synthesis for expressing attentiveness in conversational agents. In: *Proceedings of Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI), Satellite Workshop of ICMI 2016* (2016)
22. Pardo, J.S.: On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* **119**, 2382–2393 (2006)
23. Patoleta, R.: *Penis na krzyżu – gdzie przebiegają granice prowokacji?* <http://robertpatoleta.bloog.pl/id,5640692,title,penis-na-krzyzu-gdzie-przebiegaja-graniceprowokacji,index.html>. Accessed 15 Jan 2016
24. Pickering, M.J., Garrod, G.: Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* **27**, 169–225 (2004)
25. Pikus, S.: *An analysis of speech alignment in German dialogues between native speakers of German and Polish*. MA thesis. Institute of Ethnolinguistics, Adam Mickiewicz University, Poznań, Poland (2020)
26. Porzel, R., Baudis, M.: The Tao of CHI: towards effective human-computer interaction. In: Dumais, S., Roukos, S. (eds.) *HLT-NAACL 2004: Main Proceedings* (Boston, Massachusetts, USA, 2–7 May 2004), pp. 209–216. Association for Computational Linguistics (2004)
27. Porzel, R., Scheffler, A., Malaka, R.: How entrainment increases dialogical efficiency. In: *Proceedings of Workshop on Effective Multimodal Dialogue Interfaces*, Sydney (2006)
28. Savino, M., Lapertosa, L., Caffò, A., Refice, M.: Measuring prosodic entrainment in Italian collaborative game-based dialogues. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) *SPECOM 2016. LNCS (LNAI)*, vol. 9811, pp. 476–483. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_57
29. Sonar X1 LE. https://www.roland.fi/products/sonar_x1_le/. Accessed 11 Mar 2017
30. van Engen, K.J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M., Bradlow, A.R.: The wild-cat corpus of native-and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Lang. Speech* **53**(4), 510–540 (2010)
31. Ward, A., Litman, D.: Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In: *Proceedings of the SLATE Workshop on Speech and Language Technology in Education* (2007)