# Speech Corpus Creation for Automatic Analysis of Phonetic Convergence

*Grażyna Demenko, Jolanta Bachan, Agnieszka Wagner, Piotr Wyroślak*

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

**Abstract:** The paper presents the preliminary results underlying an ongoing project which aims at analysis and objective evaluation of phonetic convergence in spoken dialogues in human-human and human-machine interactions. The analyses planned to be carried out in the scope of the project will serve for (1) extracting phonetic features which can be mapped onto synthetic signal, (2) creating dialogue models applicable in a human-machine interaction and (3) practical evaluation of the convergence. The paper discusses the phenomenon of phonetic convergence, the specifications for corpus and database creation and labelling, the trial recordings and preliminary remarks on measuring lexical convergence. In the end, significance and application domains of the knowledge, methods and tools developed in the project are presented.

## 1 Introduction

Phonetic convergence can be defined as an increase in segmental and suprasegmental similarities between two speakers [1, 18]. The research on this phenomenon has its origin in the Communication Accommodation Theory (CAT) that was established in the 1970s [8, 9]. 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. The features that undergo accommodation include lexical, syntactic, prosodic, gestural and postural features, as well as turn-taking behaviour [20]. 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.

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.

We have designed our ongoing project "*Automatic analysis of phonetic convergence in speech technology systems"* to contribute to quantitative modelling of phonetic convergence in human-human and human-computer communication and to application of the convergence models in a simulated SDS environment. The main objective of the project is the analysis and objective evaluation of phonetic convergence in human-to-human and human-to-machine conversation. In the acoustic analyses, the following features are taken into account: speaking rate [24], fundamental frequency and amplitude contours [10], voice onset time (VOT) [21], utterance duration and rate [17], perceptual similarity of pronunciation [18] and spectral features. For the purpose of the analyses and investigation of the phonetic convergence phenomena, a large speech corpus including spontaneous dialogues is being created. Interactions between native speakers (Polish, German) as well as between native and non-native speakers (L1 Polish – L2 German and L1 German – L2 Polish) are being recorded. The speech corpus is being annotated at various levels (segmental, suprasegmental) and accounting for various factors (linguistic, non-linguistic, paralinguistic) and appropriate methodology for the purpose of analysing accommodation at various levels/in various domains is being formulated. On this basis a quantitative model of accommodation of segmental and suprasegmental features and a dialogue model will be proposed for implementation in an SDS environment. The model will be objectively evaluated using HMM-based speech synthesis in a Wizard-of-Oz experiment. Another objective of the project will be related to phonetic convergence in human-computer interaction – the project will examine whether users' speech converges systematically with the text-to-speech (TTS) "interlocutor". The main contribution of the project will consist in providing a quantitative description of phonetic convergence in human-human and human-computer interaction and demonstration of the implementation of accommodating behaviour in a simulated (Wizard-of-OZ) SDS environment by means of HMM-based speech synthesis.

The project addresses the current research issues in the area of conversational speech modelling and development of next-generation conversational interfaces for speech technology applications. The hypotheses tested in the project concern differences in consistency and degree of convergence between speakers (e.g. depending on speaker's gender, conversational role, language aptitude and phonetic talent), features (certain segmental and suprasegmental features will be adapted more easily and with a better outcome than others), magnitude and persistence of the convergence effects, the speed and bi-directionality of the adaptation and the re-adaptation (e.g. when speaker is introduced to a new interlocutor). As concerns spoken dialogues in human-human communication, it is expected to observe that speakers generally converge to their dialogue partners both when they communicate in their native speech and when one of the interlocutors is a non-native speaker. It is also expected to find a number of factors mediating the degree, consistency, magnitude and persistence of the convergence effects (e.g. language aptitude in a conversation in non-native speech, early and a late point of the dialogue). In a simulated SDS application using HMM-based speech synthesis, where Polish and German native and non-native speakers will be interacting with Polish and German TTS systems, it is expected to find adaptation of a variety of acoustic-prosodic features of user's speech towards a computer "interlocutor's" (TTS). Figure 1 shows the workflow of the tasks in the project. So far, the scenarios for corpus recordings were created and one trial recording and pilot annotation were carried out, along with a preliminary analysis of lexical convergence.

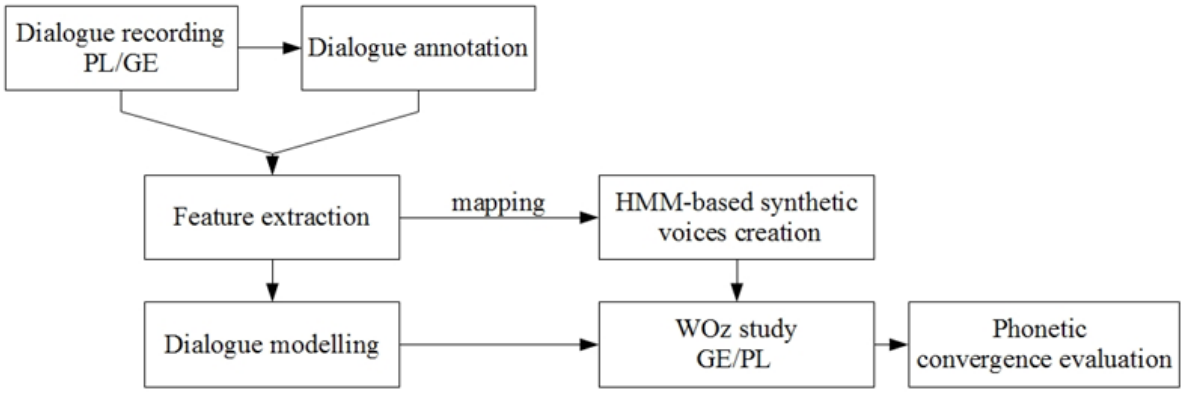



**Figure 1** - The workflow of the tasks in the project

The final result of the project will be a quantitative model of phonetic convergence in spoken dialogue, dialogue models applicable in a human-computer interaction, new synthetic voices for HMM-based speech synthesis and implications for the design of conversational interfaces in speech technology.

## 2 Specifications for corpus design, creation and annotation

### 2.1 Subjects

In the initial phase of the project, specifications for the creation of a spoken dialogue corpus and speech data acquisition method have been defined. The spoken dialogue corpora will include interactions between native speakers of Polish, native speakers of German and interactions between a native and a non-native speaker. Importantly, one goal of the project is to investigate convergence phenomena also in non-native language environments, because currently our knowledge in this field is rather scarce and only few studies concentrated on accommodation in L2 communication [14, 21, 29].

It is planned to record the following pairs of speakers:

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

It is planned to record pairs of young students, possibly people who know each other in order to assure the best possible conditions for speaker alignment [1, 3]. For each of these 4 groups, 10 pairs (20 people) are going to be recorded.

For each speaker, we will collect information such as name, sex, age, height, weight, education, possibly profession, languages spoken and proficiency level in L2 German or Polish. This kind of metadata will be stored together with the annotations in the database (see sec. 2.4).

### 2.2 Levels of annotation characterising speakers

The selection of annotation levels in the project is based on the delineation of three classes of features: *physiologically-conditioned acoustic*, *phonetically-conditioned acoustic* and *grammatical*. *Physiologically-conditioned acoustic* features pertain to fundamental frequency, loudness, timbre and accuracy of articulation. Acoustic phenomena considered to be more linguistically motivated (such as prosody, rhythm and suprasegmental phonetic interactions) fall into the *phonetically-conditioned acoustic* group. The last set of features is related to annotation aimed at capturing the information from semantic, lexical and morphosyntactic levels (for more detail see [7]). Including such description in the research may aid both the

broadly conceived study of convergence in human communication and the assessment of the relevance of nonphonetic features to the creation of effective linguistic models for application in human language technology.

### 2.3 Scenarios for the recording sessions

The corpus is to consist of recordings from sessions conducted according to three scenarios: neutral, expressive and controlled. The neutral scenario consists of a diapix task [27], a maptask and a task aimed at reaching an agreement between conversational partners (subjects are to decide which items from a given list prove most useful during a stay on a desert island). In the first part of the expressive scenario, participants are involved in conversations on controversial works of modern art: The aim of the first task is to discuss three photographs of an artistic installation, Figure 2a. Thereafter, subjects are asked to imitate a dialogue between an enthusiast of provocation in modern art and a person with a more conservative view, Figure 2b. The remaining parts of the expressive set consists of tasks based on conversations between information seekers and information providers ("party-goer" and tourist information centre assistant; prospective client and travel agent; journalist and museum visitor). In the controlled scenario pairs of participants are supposed to read an interview with a popular musician (person A reading the parts by interviewer and person B those by the interviewee). Afterwards, they are asked to listen to the recording of their own turns and read the corresponding parts so as to fit in the dialogue. The motivation for this task is to test how speakers align to their own speech.

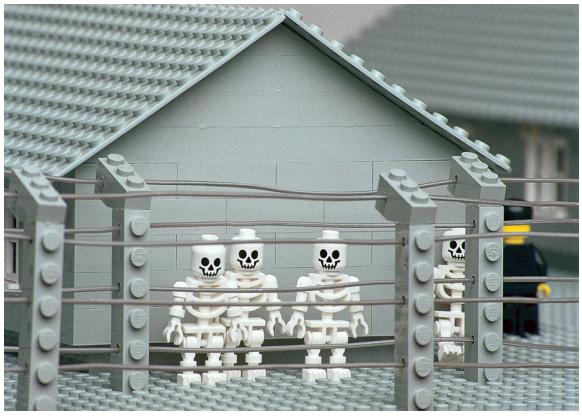
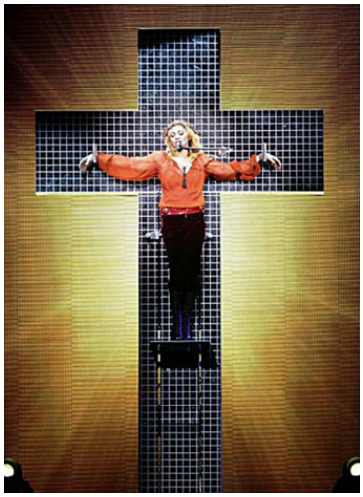
a)  b) 

**Figure 2 –** a) "Lego. Concentration Camp." by Zbigniew Libera [15]; b) Madonna on the cross [19]

### 2.4 Labelling

For the purpose of speech corpus annotation and analysis, the segmentation and annotation tools *Praat* [5]*, Transcriber* [26], *Salian* [25] and *Annotation Pro* [11] are applied. *Praat* and *Transcriber* are typical annotation programs which offer different functionalities and therefore are efficient for different tasks. *Annotation Pro* is a recently developed tool whose functionalities include among others multilayer annotation of spoken utterances with categorical and continuous features (e.g. inserted via a graphical interface), speech segmentation at the phoneme, syllable and word level [25], analysis of timing properties of utterances using the *Annotation Pro* plugins for Time Group Analysis [13] and rhythm metrics, and extraction of the annotation and analysis results for the needs of their further processing. *Annotation Pro* enables also efficient design and conduction of perception tests and therefore can be used for perceptual assessment of the presence of convergence and its character in terms of degree, magnitude and persistence, and the speed and bi-directionality of the adaptation and the re-adaptation. For the purpose of annotation management and storing,

we plan to design a relational database using Microsoft's SQL Server client/server database engine as previously used in the Jurisdict project [12]. The implemented database model will reflect the structure of the corpus and will be integrated with *Annotation Pro* and possibly other tools that will be applied at different stages of corpus annotation and analysis.

The speech corpus will be provided with the following annotation tiers: orthographic transcription, phonetic transcription, dialogue acts [6], fluent vs. disfluent phases of dialogue, prosodic structures of utterances (on the linguistic and non-linguistic levels). As concerns prosody, we have adopted the labelling scheme developed within a project on a Polish ASR system [2]. In the description of prosodic phenomena, the following factors are taken into consideration: two levels of syllable prominence – *strong* and *weak-medium* prominence enhancement (marked by numbers 1, 2, 3), and three levels of prosodic phrase boundary strength – the *weak*, *medium* and *strong* boundaries (marked by slashes /, //, ///). In addition, elements of discourse with a high impact on the prosodic structure of speech are taken into account. An excerpt of a trial dialogue with annotation is presented on Figure 3.

The annotation of prominence and phrase boundaries is 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 in places where syntactically and semantically such a boundary occurs, but the acoustic cues are very subtle, and (2) labels indicating "ungrammatical" boundaries 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.

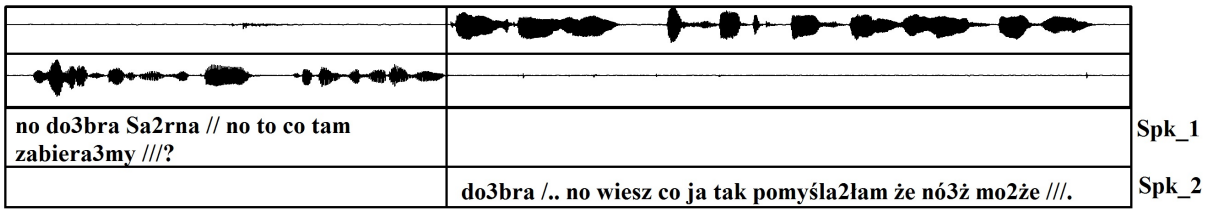



**Figure 3 –** Excerpt from the annotation of trial recordings

# 3 Trial recordings

## 3.1 Recording settings

The first recording session (one pair of speakers) was carried out in December 2015 and lasted 100 min. From the designed scenarios, 54 min of dialogues were obtained. The recordings were carried out in a professional recording studio and were acoustically and visually separated. The speakers could hear each other over headphones, but could not see each other. For the recordings, 4 microphones were used: 2 overhead microphones (DPA 4066 Omnidirectional Headset Microphone) and 2 stationary microphones (diaphragm microphone, Neumann TLM 103). This setup provided 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 [23] and Roland Studio Capture hardware was the audio interface employed.

## 3.2 Measuring lexical convergence in a Polish dialogue – preliminary analysis

Apart from the analyses of convergence for the phonetic features, the relevance of the lexical data may be evaluated. However, not until effective measures of lexical convergence in dialogue are available, can the predictive power of lexical data in linguistic models for human language technology be assessed. The existing measures are based on a degree of overlap in the vocabulary used, and differ in several respects. A salient difference is in the parts of the dialogue in which the overlap is tested: both counts of word repetition in moving windows of

N turns [28] and exploitation of the data from overall frequency lists of conversations [16] have been proposed. An even more fundamental question concerns the type of lexical unit taken as the domain of the analysis (e.g. lemmas, words derived from the same stem). There is also uncertainty about which classes of lexical items could be taken into consideration as most revealing. Ward and Litman [28] indicate that, ideally, the items with no equivalent available should be excluded so as to enhance the precision of the study. Nenkova et al. [16] focus on the classes delineated from the most frequent words, for example, affirmative cue words.

A preliminary qualitative analysis on material from corpora assembled may encourage investigation of such a class, cf. the distribution of the Polish word *no* (En. *well*, *so*, *yeah*) in the following excerpt:

Speaker1: *nóż no tak no ja bym wziął w ogóle zapalniczkę i zapałki wszystko co z ogniem nie?*
(En. *knife...well yeah…well... I would take...in general...a lighter and matches...everything that goes with fire...so?*)
Speaker2: *no no wiadomo że to na początek nie no tam to te*
(En. *well...well….I know, for the beginning...well... so...there...these*)

It remains unclear whether existing measures of lexical convergence, evaluated on the material of English dialogues, may be successfully applied to highly-inflected languages such as Polish. The questions of stemming and lemmatisation become even more acute – while the approach relying solely on type repetition may be very prone to the data sparsity problem, morphologically encoded information may be of relevance to the models based on the study of communicative alignment. It might be hypothesised that taking the potential interdependence between lexical and morphosyntactic features into consideration may prove to be effective in providing effective measures of convergence on these levels. Both the need for empirical evidence for such claims and the potential exploitation of the findings can benefit from the use of novel approaches to multivariate statistical analyses in linguistics and natural language processing.

## 4 Significance of proposed research

The proposed research is of key importance for the speech communication sciences and technologies. It deals with the fundamental unsolved problem of phonetic convergence which is responsible for a number of theoretical and practical difficulties, e.g. in speech analysis and in the implementation of speech synthesis and speech recognition. The project will also contribute to progress in the processing of paralinguistic and non-linguistic information in speech, which is highly significant for various speech technology applications such as speaker identification and characterization, speech recognition and synthesis, and human-machine communication in general.

Aside from the scientific relevance, the project has significant application value, especially for telecommunication systems, where speech naturalness is highly desired, as well as in the area of automation of office work, speaking databases, personalised speech synthesisers for the disabled and application in speaking websites. Technological benefits that can be expected from the project results include also the development of voice user interface enabling human-computer interaction through voice/speech platform.

A better understanding of the accommodation phenomena related to various acoustic-prosodic, temporal and spectral features may directly improve the performance of current SDS technology leading to smoothness of conversational dialogue based on temporal accommodation, positive evaluation of the overall interaction by the user, improvement of prosodic models for synthesised speech, and/or improving performance of Automatic Speech Recognition (ASR) by exploiting user adaptation to the system voice [4]. Among interacting

talkers, phonetic convergence might contribute to mutual comprehension and/or rapport through a decrease in social distance [22].

Finally, while most previous studies investigated phonetic convergence in a native language environment, we believe that the current project may contribute to the understanding of various aspects of second language face-to-face communication by exploring possible underpinnings of phonetic convergence in an L2-environment.

## 5 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."*

## References

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

[2] Bachan, J., A. Wagner, K. Klessa and G. Demenko: Consistency of Prosodic Annotation of Spontaneous Speech for Technology Needs. In: Vetulani, Z. and J. Mariani (Eds.): Proceedings of the 7th Language & Technology Conference, Poznań, Poland, 2015.

[3] Batliner, A., S. Steidl, Ch. Hacker and E. Nöth.: *Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech*. User Modelling and User-Adapted Interaction – The Journal of Personalization Research 18, 2008, pp. 175 – 206.

[4] Bell, L., J. Gustafson and M. Heldner: *Prosodic adaptation in human-computer interaction*. 15th International Congress of Phonetic Sciences, 2003.

[5] Boersma, P. and D. Weenink: Praat: doing phonetics by computer. Version 6.0.11 Available: http://www.fon.hum.uva.nl/praat/ [Accessed: January 18, 2016]

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

[7] Demenko, G.: *Korpusowe badania języka mówionego.* [En. *Corpus studies of spoken language*] Warszawa: Akademicka Oficyna Wydawnicza EXIT 2015.

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

[9] 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. Cambridge: Cambridge University Press 1991, pp. 1 – 68.

[10] Gregory, S. W.: *Analysis of fundamental frequency reveals covariation in interview partner's speech*. Journal of Nonverbal Behavior 14, 1990, pp. 237 – 251.

[11] Klessa, K.: Annotation Pro. Version 2.2.6.0. Available: http://annotationpro.org/ [Accessed: May 19, 2015]

[12] Klessa, K. and G. Demenko: Structure and Annotation of Polish LVCSR Speech Database. In: Proceedings of Interspeech Conference, Brighton, UK, 2009.

[13] Klessa, K. and D. Gibbon: Annotation Pro + TGA: automation of speech timing analysis. In: Proceedings of the 9th Language Resources and Evaluation Conference, Reykjavik, Iceland, 2014.

[14] Lewandowski, N.: *Talent in nonnative phonetic convergence*. Ph.D. Thesis Universität Stuttgart, 2012. Available: http://elib.uni-stuttgart.de/opus/volltexte/2012/7402/ [Accessed: March 5, 2013]

[15] Libera, Z.: Lego. Obóz Koncentracyjny, 1996. [Photography of the work from the resources of Galeria Raster] Available: http://culture.pl/sites/default/files/

images/imported/sztuki%20wizualne/lego2.jpg [Accessed: January 15, 2016]
[16] Nenkova, A.; A. Gravano and J. Hirschberg: High frequency word entrainment in spoken dialogue. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, 2008.
[17] Oviatt, S., C. Darves and R. Coulston: *Toward Adaptive Conversational Interfaces: Modeling Speech Convergence With Animated Personas*. ACM Transactions on Computer-Human Interaction 11, 2004, pp. 300 – 328.
[18] Pardo, J. S.: *On phonetic convergence during conversational interaction*. Journal of the Acoustical Society of America 119, 2006, pp. 2382 – 2393.
[19] Patoleta, R.: Penis na krzyżu – gdzie przebiegają granice prowokacji? Available: http://robertpatoleta.bloog.pl/id,5640692,title,penis-na-krzyzu-gdzie-przebiegaja-granice-prowokacji,index.html [Accessed 15 January, 2016]
[20] Pickering, M. J. and S. Garrod: *Toward a mechanistic psychology of dialogue*. Behavioral and Brain Sciences 27, 2004, pp. 169 – 225.
[21] Sancier, M. L. and C. A. Fowler: *Gestural drift in a bilingual speaker of Brazilian Portuguese and English*. Journal of Phonetics 25, 1997, pp. 421 – 436.
[22] Shepard, C. A., H. Giles and B. A. Le Poire: Communication accommodation theory. In: Robinson, W. P. and H. Giles (Eds.): The New Handbook of Language and Social Psychology. New York: Wiley 2001, pp. 33 – 56.
[23] Sonar X1 LE Cakewalk Production Software – Copyright ©, Roland Corporation – Available: http://www.roland.com/products/sonar_x1_le/ [Accessed January 15, 2016]
[24] Street, R. L. Jr.: *Speech convergence and speech evaluation in fact-finding interviews*. Human Communication Research 11, 1984, pp. 139 – 169.
[25] Szymański, M. and S. Grocholewski: Transcription-based automatic segmentation of speech. In: Proceedings of 2nd Language & Technology Conference, Poznan, Poland, 2005, pp. 11 – 15.
[26] Transcriber – Copyright © 1998-2008, DGA. Available: http://trans.sourceforge.net/ [Accessed: January 15, 2016]
[27] 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, December 2010, pp. 510 – 540.
[28] Ward, A. and D. Litman: 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.
[29] Zuengler, J.: Accommodation in native-nonnative interactions: Going beyond the "what" to the "why" in second language research. In: Giles, H., J. Coupland and N. Coupland (Eds.): Contexts of Accommodation: Developments in Applied Sociolinguistics. Cambridge: Cambridge University Press 1991, pp. 223 – 244.