A Corpus of Spontaneous Multi-party Conversation in Bosnian Serbo-Croatian and British English

Emina Kurtić[1,2], Bill Wells[2], Guy J. Brown[1], Timothy Kempton[1], Ahmet Aker[1]

[1]Department of Computer Science,

[2]Department of Human Communication Sciences
University of Sheffield, UK
e.kurtic@sheffield.ac.uk, b.wells@sheffield.ac.uk, g.brown@dcs.shef.ac.uk, t.kempton@dcs.shef.ac.uk, a.aker@dcs.shef.ac.uk

Abstract

In this paper we present a corpus of audio and video recordings of spontaneous, face-to-face multi-party conversation in two languages. Freely available high quality recordings of mundane, non-institutional, multi-party talk are still sparse, and this corpus aims to contribute valuable data suitable for study of multiple aspects of spoken interaction. In particular, it constitutes a unique resource for spoken Bosnian Serbo-Croatian (BSC), an under-resourced language with no spoken resources available at present. The corpus consists of just over 3 hours of free conversation in each of the target languages, BSC and British English (BE). The audio recordings have been made on separate channels using head-set microphones, as well as using a microphone array, containing 8 omni-directional microphones. The data has been segmented and transcribed using segmentation notions and transcription conventions developed from those of the conversation analysis research tradition. Furthermore, the transcriptions have been automatically aligned with the audio at the word and phone level, using the method of forced alignment. In this paper we describe the procedures behind the corpus creation and present the main features of the corpus for the study of conversation.

Keywords: multi-modal corpus, naturalistic conversation, turn taking

1. Introduction

Mechanisms that underlie spoken human interaction have been extensively studied both to understand the phenomenon for its own sake and, more recently, to develop technological applications in which humans communicate with machines over speech interfaces. People talk in a variety of settings: phone conversations, face-to-face or virtual meetings, in different social situations, etc. Some of these take place within agreed social circumstances which can influence the kind of interaction. For example, in institutional conversations such as meetings the manner of spoken exchange can be determined by the meeting agenda, the presence of a person chairing the meeting, or simply the fact that all participants know the purpose of the meeting (Sacks et al. 1974). These factors are absent in spontaneous, mundane, every-day conversation, in which people get together "just to talk". It has been argued that this latter type of conversation is the best environment for understanding the organisation of human spoken interaction (Levinson 2006).

Studying mundane conversation has a longstanding tradition within Conversation Analysis (CA) and the linguistic discipline of interactional linguistics (IL), which developed from it. By detailed investigation of conversational sequences, CA and IL researchers were able to uncover the types of social action the conversations contain, as well as to describe how participants use language to conduct these. Recent research in IL in particular (Kurtic et al. (2010), Gorisch et al. (2012)), suggests that applying these types of analyses to large data sets of conversational phenomena can offer an integrated picture of the use of language and non-verbal cues in the realization of conversational acts, which analyses of more limited numbers of conversational sequences could not discover. The analysis of linguistic detail in large data sets calls for automatic processing of spoken and written language, as well as methods for the analysis of non-verbal cues in conversation.

A precondition for application of these methods is the availability of high quality multi-modal recordings of human conversations. Such corpora should fulfil several criteria:

(i) They should contain naturally occurring, face-to-face, non-institutional talk.

(ii) The quality of audio recording should be sufficient to allow reliable automatic analysis of speech related features (e.g. fundamental frequency, speech intensity). This should be the case even in situations where speakers overlap. Currently, this is best achieved by recording speakers on separate audio channels.

(iii) Minimally, video recording should be provided along with audio for analysis of non-verbal features of conversation.

(iv) There should be sufficient amounts of both two-party and multi-party conversational data, since participants in multi-party conversations draw on different strategies of conversation management from those in two-party ones. For example, the assignment of speaker-addressee roles in two-party conversation is exhausted by the number of speakers, whereas in multi-party conversation, there may be other roles, like that of non-addressed listener etc., which need to be made clear between participants (Levinson 2006).

(v) Corpora should be collected in different languages, to allow for investigation of cross-cultural and comparative linguistic investigations of interactional practices.

Although recent years have seen publication of an increasing number of conversational corpora, few of them fulfil all these criteria. D64 (Oertel et al. 2010) is an example of a multi-party, face-to-face conversation recording in English, which contains a portion of casual conversation in addition to work-related talk. Furthermore, some examples of two-party recordings of spontaneous conversation are Spontal (Edlund et al. 2010) for Swedish, SpontalN (Sikveland et al. 2010) for Norwegian, the Nijmegen Corpus of Casual French and Spanish (Torreira et al. 2010 and Torreira & Ernestus 2010, respectively).

In this paper we describe a corpus of spontaneous, casual, face-to-face conversation between friends, recorded in two languages, Bosnian Serbo-Croatian (BSC) and British English (BE). With this corpus we aim to contribute further data that fulfils the above criteria. The corpus will be published as a freely available resource of naturally occurring talk, suitable for investigation of a variety of research questions related to spoken interaction. Recording under similar conditions in both languages permits comparative, cross-linguistic research on multi-modal aspects of spoken interaction, as well as research on each of the target languages separately.

The present corpus is the first freely available resource of spoken BSC. Although some aspects of BSC have been well studied (for example, its word-accent system), empirical, corpus-based studies of spontaneous spoken language do not currently exist. The present corpus is thus a unique resource for such study, and will contribute to the development of automatic methods for processing spoken and written BSC.

2. Data collection

2.1 Speakers

The participants were native speakers of each language, three female and one male.

The four native speakers of BSC come from the city of Tuzla in the north-eastern region of Bosnia and Herzegowina. At the time of the recording the speakers were final year undergraduate students or PhD students and teaching assistants at the University of Tuzla. The speakers were friends well acquainted with each other through university life.

The four BE speakers were a group of friends, all students at the University of Sheffield at the time of the recording. The female speakers were from London and the male speaker from Sheffield.

2.2 Recording set-up

The recordings were made in university rooms familiar to all participants (Figure 1). The recording setting mirrored an informal meeting for social talk: food and drink were provided on the table and the participants were eating and drinking during the recordings; no instructions were given as to possible topics of the conversation, and the participants were free to get up and walk around during recordings although they did not make use of this. The recordings were made over two days, with two subsequent recording sessions of approximately one hour per day.

Digital audio recordings were made on a MacBook Pro laptop computer, which controlled a MOTU 8Pre FireWire audio interface. Each of the four participants was recorded via an individual headset microphone (Sennheiser ME 3-N cardioid headset, powered from the MOTU interface via MZA-990P phantom power adaptors). In addition, an omnidirectional audio recording was made via a pressure zone (PZM) microphone on a fifth audio channel. Sound recordings were made using MOTU AudioDesk software at a sample rate of 44.1 kHz with 16 bit resolution (BSC) and 48 kHz (BE), with 24 bit resolution (BE). This forms the core of the data available in both languages.

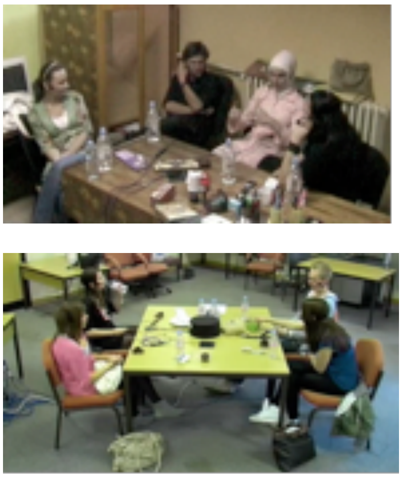

Figure 1: Recording setting for BSC (top) and BE (bottom).

BE conversations were additionally recorded using a microphone array for use in speech recognition and speaker diarisation (Marino and Hain, 2011). The microphone array contained 8 microphones and was placed in the middle of the table around which participants were seated.

Digital video recordings were made with a Canon (MV600) camera. An additional camera (Canon XM2) was used in the BE recordings, to capture all participants from an additional angle. The cameras were positioned on the elevated surface in a corner of the recording rooms in order to capture as much of participant's body movements as possible given the constraints of the rooms.

3. Segmentation and Transcription

All recordings were segmented and transcribed by trained transcribers. The following sections describe the decisions made in the segmentation and transcription process.

3.1 Segmentation

The spoken speech stream can be segmented into units of different size and type. Previous reports on segmentation of conversational speech suggest that the choice of segmentation unit largely depends on the main purpose the corpus is created for. For example, for automatic speech recognition, it is important that units

Speaker	TCUs
enM2	it's all we talk about.
	(0.51)
enM2	you can't tell who goes to the toilet.
enM2	whether it's him or her.
	(0.24)
enM2	because they wear like this red hoody.
enM2	and put hood up.
enM2	and walk past like.
	(0.07)
enM2	just they creep past to toilet.

Table 1: Segmentation of speaker's *enM2* extended speaking turn into TCUs. Each TCU is in a separate row of the table. The numbers in brackets indicate the pause duration (in ms) between TCUs.

Non-speech sounds: inbreath, outbreath, voice onset, clear throat, lip smack, click, bilabial trill, silent laughter, loud laugther, giggle, outbreath laughter, inbreath laughter, room noise, writing, door slam, mobile, silence, unidentifiable noise, channel noise, sniff, blowing nose, cough, hiccough, sneeze, whistle, yawn, prolonged sound, prolonged vowel, unidentifiable vocalisation, other

Table 2: Non-speech sounds

have a manageable size, to avoid the data sparsity problem in recogniser training, which can arise in the case of longer sequences. Therefore, for example, the ICSI corpus has been segmented into time bins, which are "practical units, rather than theory-relevant" (Edwards 2004). In dialogue act annotation, on the other hand, it is important that a unit carries potentially multiple pragmatic actions. So-called "functional segments" (Bunt 2011) have been proposed for this purpose.

To understand how conversation is progressed by its participants, we believe that it is essential that units of talk are identified according to how participants in conversation themselves understand the organization of talk. We therefore endorse the notion of the turn constructional unit (TCU) proposed by the turn-taking model of Sacks et al. (1974).

TCUs can be understood as minimal potential speaker turns. They are turn-taking units at the end of which speaker change becomes relevant and legitimate, but does not need to happen. To prevent speaker change from happening at the end of a TCU, participants will mostly use turn-holding devices like a particular pitch contour (Duncan (1972), Cutler & Pearson (1986), Wells & MacFarlane (1998)), a rush through (Walker 2010) or an abrupt joint (Local & Walker 2004) around possible TCU completion points. Likewise, participants have means of signaling the end of a TCU, as identified by previous research. Ford & Thompson (1996) for example conclude that a TCU is a unit which is syntactically complete, implements a recognizable action (i.e. is pragmatically complete) and is also intonationally coherent.

An example of segmentation of talk into TCUs is given in Table 1. This extract is taken from the BE corpus and shows speaker enM2's extended turn with no interference from other speakers segmented into TCUs. Segmenting conversation into TCUs has the advantage that it fulfills both the above requirements. The TCU is the minimal constituent of a turn and so is typically short, while still carrying a complete pragmatic meaning. Most importantly, however, the TCU is the building block that participants themselves use when constructing their talk.

3.2 Transcription

For transcription, the transcription and annotation tool ELAN (Wittenburg et al. 2006) was used. Each transcript contains four information tiers: (i) orthographic transcription, (ii) annotation of non-speech sounds, (iii) uncertain and (iv) comments.

On the orthographic transcription tier, each TCU was transcribed using the standard orthography of each language. Conventions were devised to transcribe unintelligible words in an approximate spelling that reflects their pronunciation and in addition to mark them as uncertain on the uncertain tier. Conversation is rich in interjections, fillers and response tokens, like uh, uhhuh, mmm, hmm, etc., many of which lack an orthographic equivalent. These were transcribed according to the conventions used in CA, which aim to reflect the way they are heard.

The tier *non-speech* sounds contains information about non-linguistic sounds, which nevertheless can have a conversational function. The list of non-speech sounds included is given in Table 2.

The *comments* tier was made available so that transcribers could record additional notes during transcription; however, it was rarely used by the transcribers.

3.3 Automatic alignment of transcripts at the word and phone level

Many research tasks, in particular those involving the study of phonetic detail in conversation, require segmentation of speech below the TCU level. For this reason, we provide segmentation of the corpus at the word and phone level as well. Manual segmentation is a very labour intensive task, which is also prone to errors and variation. It has been repeatedly argued (e.g. Sikveland et al. 2010) that automatic segmentation, apart from being time and labour efficient, also has the advantage that the errors are predictable. Therefore, our segmentations at the phone and word levels are created automatically. For the automatic alignment at phone level we use forced alignment, in which a speech recogniser is used to identify the start and end times of phones and words based on the transcript and the speech signal. In the following sections we describe and evaluate the phone level alignment

Building a speech recogniser requires a pronunciation dictionary for the language and acoustic models trained for the specific language and type of the

Phone set of the recogniser	20ms Error
Czech	53%
Russian	51%
Hungarian	59%
American English (TIMIT)	57%

Table 3: Results of BSC forced alignment. 20ms Error indicates the proportion of boundaries placed more than 20ms away from the ground truth boundary.

data (in this case conversational speech). For BE we use the pronunciation dictionary and acoustic models trained on the AMI corpus (Hain et al. 2007). For evaluation we compared the automatically forced aligned data with a ground truth phone level segmentation generated by correcting a portion of 100 randomly selected TCUs (1434 phone boundaries). The performance of the recogniser was evaluated with software provided by Hossom (2009). This gives the standard forced alignment evaluation error: the proportion of boundaries placed more than 20ms away from the ground truth boundary (20ms Error).

In the evaluation of BE data alignment the 20ms Error was 35%. In a qualitative error assessment we found that misalignment was mainly found in cases where laughter or outbreaths were overlaid on speech or in regions of whispered or creaky voice. Also turn final word lengthening, often associated with creaky voice was frequently misaligned, as well as non-standard pronunciations such as found in acronyms for example.

For BSC no pronunciation dictionary or acoustic models trained for conversational speech are available at present. The success of building these resources from our data is questionable given that the amount of available recordings is considered small for this purpose. We therefore consider the method of cross-language forced alignment described in Kempton et al. (2011).

Cross-language forced alignment enables an under-resourced language to be forced aligned using a phone recogniser that was trained on a different language. Its main aim is to give a good initial alignment on small amounts of challenging data in the under-resourced language. Phone labels are automatically mapped to the closest phone of the best available recogniser. We evaluate the cross-language forced alignment performed by recognisers trained for Czech, Russian, Hungarian and American English (Schwarz 2009). The cross-language forced alignment is evaluated by comparison to human generated ground truth phone level segmentation, which was performed on 124 randomly selected TCUs (1468 phone boundaries). The results are shown in Table 3.

The results indicate that the Russian phone recogniser performs best when used for BSC forced alignment. It significantly outperforms the Hungarian and AE recognisers ($p<0.01$) as indicated by the related samples Wilcoxon signed rank test used with the Bonferroni correction. It also outperforms the Czech recogniser, however, this difference is not significant. Given these results, the full 3 hours of the BSC corpus was automatically aligned with the Russian phone recogniser.

The alignment results were further evaluated qualitatively to establish the error sources and identify the potential for further improvement. The qualitative evaluation showed that an appreciable number of errors is related to spontaneous speech phenomena, like creaky voice, quick and silent articulations at TCU beginnings and ends, laughter and outbreath overlaid on speech and loud inbreaths. Also the vocalization without standard orthography, like ‘uh’ proved problematic. A further source of misalignments were shortening the vowels in vowel-nasal/plosive/glide transitions and taking only the duration of the closure to be a plosive, although there is usually the release with some aspiration which was frequently left out. The errors seemed to be more frequent in false starts, or short TCUs, and absent in longer stretches of ‘grammatically’ correct talk. A test of correlation between TCU length and forced alignment error shows a significant negative correlation $\rho=-0.52$ ($p<0.001$), indicating that longer TCUs are indeed aligned more accurately.

These observations indicate that incorporating knowledge about vowel duration and closure duration in plosives along with adjusting the Russian phone recogniser to spontaneous talk phenomena would substantially improve the accuracy of the forced aligned data for BSC. The results also indicate that cross-language is a viable option for segmentation of BSC for which currently few resources exist.

4. Availability

The corpus will be made available under the Shareware Creative Commons Licence for free use for research purposes. It will be accessible from July 2012 via a Java-based web site which allows the corpus to be browsed and searched for annotated items of interest (e.g. all tokens of “Yeah”, “like”, “Uh”, all inbreaths, etc.). The search engine also provides facilities for searching on particular audio channels, and for identifying overlapping speech involving two or more participants. The audio and video clips returned by a search can be downloaded individually, or together in the form of a ZIP archive. In this way, the corpus will be more accessible to its users who will be able to gain a quick insight into corpus features without having to download it first and browse it using offline tools.

5. Conclusion and Future Work

We presented a corpus of multi-modal recordings of spontaneous, face-to-face, multi-party conversation in Bosnian Serbo-Croatian and British English. The corpus is aimed as a general-purpose resource for the study of spoken interaction. It consists of 3 hours of free conversation in each of the target languages. The audio recordings have been made on separate channels using head-set microphones, as well as using a microphone array, containing 8 omni-directional microphones. The data has been segmented and transcribed using the segmentation notions and transcription conventions derived from those developed within the conversation analysis research. The transcriptions have been automatically aligned with audio at the word and phone level using forced alignment.

Our current work on the corpus includes improvement of the forced-alignment accuracy with further training iterations (van Niekerk & Barnard 2009) and refinements based on phonetic knowledge (Peddinti & Prahallad 2011). We are also exploring methods of low-cost, non-intrusive motion capture, which will enable us to supplement video recordings with recordings of head-movements, gestures and other motions that participants use in conversation. In future work, we plan to include a variety of less-studied languages into the corpus and thus create a rich resource for cross-cultural and comparative linguistic study of conversation.

6. Acknowledgements

This work was supported by the UK Arts and Humanities Research Council (AHRC). We are grateful to Matt Gibson and Thomas Hain for advice and assistance with forced alignment for British English. We also thank all who took part in or facilitated the creation of the corpus.

7. References

Bunt, H. (2011). Multifunctionality in Dialogue. *Computer, Speech and Language,* 25, pp. 225-245.

Cutler, A. & Pearson, M. (1986), On the analysis of turn-taking cues. In C. Johns-Lewis, ed., *Intonation in Discourse*, Croom Helm, London, pp. 139–155.

Duncan, S. (1972), Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23(2), pp. 283–292.

Edwards, J. (2004). The ICSI Meeting Corpus: Transcription methods. ICSI Meeting Corpus Documentation. ICSI, Berkley, CA.

Ford, C. & Thompson, S. (1996), Interactional units in conversation: Syntactic, intonational and pragmatic resources for management of turns, in E. Ochs, E. Schegloff & S. Thompson (Eds.), *Interaction and Grammar*, Cambridge University Press, Cambridge, UK, pp. 134–184.

Gorisch, J., Wells, B., & Brown, G. J. (2012). Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation. *Language and Speech,* 55(1), pp. 57–76.

Hain, T., Burget, L., Dines, J., Garau, G.,Wan, V., Karafiat, M., Vepa, J. & Lincoln, M. (2007). The AMI system for the transcription of speech in meetings. In *Proceedings of* ICASSP07, Vol. 7. Honolulu, Hawai'i, USA. pp. 357-360.

Hossom, J.P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51 (4), pp. 352-368.

Kempton, T., Moore, R. and Hain, T. (2011). Cross-language phone recognition when the target language phoneme inventory is not known. In *Proceedings of Interspeech* 2011, Florence, Italy.

Kurtić, E., Brown, G. J. and Wells, B. (2010). Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration. In *Proceedings of Interspeech 2010*, Mahukari, Japan. pp. 2550-2553.

Levinson, S. C. (2006). On the human "interaction engine". In N. J. Enfield, & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction*. Oxford: Berg. pp. 39-69.

Local, J. & Walker, G. (2004). Abrupt-joins as a resource for the production of multi-unit, multi-action turns. *Journal of Pragmatics,* 36(8), pp. 1375–1403.

Marino, D. and Hain, T. (2011). An analysis of automatic speech recognition with multiple microphones. In *Proceedings of Interspeech 2011*, Florence, 28th-31st August.

Oertel, C., Cummins, F., Campbell, N., Edlund, J., & Wagner, P. (2010). D64: A corpus of richly recorded conversational interaction. In *Proceedings of* LREC '10. Valetta, Malta.

Peddinti, V. and Prahallad, K. (2011). Exploiting phone-class specific landmarks for refinement of segment boundaries in TTS database. In *Proceedings of Interspeech* 2011. Florence, Italy.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language,* 50, pp. 696-735.

Schwarz, P. (2009) Phoneme Recognition Based on Long Temporal Context, *Ph.D. thesis*, Brno University of Technology, 2009.

Sikveland, A., Öttl, A., Amdal, I., Ernestus, M., Svendsen, T., & Edlund, J. (2010). Spontal-N: A Corpus of Interactional Spoken Norwegian. In *Proceedings of* LREC'10. Valetta, Malta. pp. 2986-2991.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52, pp. 201-212.

Torreira, F., & Ernestus, M. (2010). The Nijmegen corpus of casual Spanish. In *Proceedings of* LREC'10. Valetta, Malta. pp. 2981-2985.

van Niekerk, D.R. and Barnard, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. In *Proceedings of Interspeech* 2009. Brighton, UK. pp. 880-883.

Walker, Gareth 2010. The phonetic constitution of a turn-holding practice. In Prosody in Interaction, Barth-Weingarten, Dagmar, Elisabeth Reber and Margret Selting (eds.), 51–72.

Wells, B. & McFarlane, S. (1998), Prosody as an interactional resource: Turn projection and overlap, *Language and Speech,* 41(3–4), pp. 265–294.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of* LREC 2006. Genoa, Italy. pp. 1556-1559.