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

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

Department of Human Communication Sciences[1] and Department of Computer Science[2], 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

Motivation

- Are the temporal and phonetic characteristics of overlapping talk similar across languages?
- To answer this question, need high quality recordings made under the same conditions in two languages.

Characteristics of the Corpus

- Collected in British English (BE) and Bosnian Serbo Croatian (BSC) to allow comparative investigations;
- High quality audio and video recordings;
- Naturally occurring, face-to-face talk;
- Non-instututional, since spoken exchange in institutional meetings can be influenced by the agenda etc. [5]
- Each participant recorded on a separate audio channel to allow reliable analysis of acoustic features.

Example from the Bosnian Corpus

```
F1  će jest, oooj
M1  pa dobro znaš kako ima
F2  cvijeće dolaze familija [tetke koje su došle-]
    flowers family is coming [aunts who came- ]
F3                          [opet to ima ne- ] znaš opet to ima neku čar
    [again it has so- ] you know it still has some appeal
M1  Imaš ljude ko- koji, ono
    there are people who- who, you know
F2  al naš šta ću ja reć ja ću reć evo ne znam vinarija crveni salon bilo gdjee-
    but you know what I'll say I'll say here I don't know the winery the
    red saloon anywhere
M1  Čekajte tamo ja ću doći
    wait there I'll come
F3  al  [nije to njima to]
    but [that's not the same for them]
F2      [ne znam čekaj eto mene] za deset minuta
    [I don't know wait I'm ] coming in ten minutes
F3  nije to njima to razumieš-
    that's not the same to them you understand-
F3  oni da im [je da to- ]
    they would like [to have it-    ]
F2            [al znam al da ] mi dolaze sad tu
    [but I know but that] they are coming now here
F3  al razumijem i tebe opet
    but I understand you as well
```

- F2 and F3 talk about the practice that family attends the public defence of theses (viva)
- F2 doesn't want anyone to come to the viva, F3 argues that it's important for the family
- They overlap each other many times in the extract

Data Collection and Recording

- Participants were native speakers of each language, three female and one male.
- Existing friendship groups (students).
- Informal setting in univeristy room (e.g., participants could eat and drink).
- 3 hours of recordings for each language.
- Digital audio recordings made on MacBook Pro, using AudioDesk software and MOTU 8Pre Firewire audio interface.
- Participants recorded with Sennheiser ME 3-N cardioid headsets.
- Omnidirectional recording with PZM microphone. BE meetings also recorded with an array of 8 microphones [4].
- Video recordings made with Canon MN600 camera. Additional Canon XM2 camera used for the BE recordings.

Recording setup for Bosnian Serbo Croatian at the University of Tuzla, Bosnia and Herzegovina (top) and setup for British English recordings at the University of Sheffield, UK (bottom).

Segmentation

- Speech was segmented into turn constructional units (TCUs) [5].
- TCU is a minimal constituent of a speaker turn; syntactically and pragmatically complete, building block used by talkers.

Transcription

- Transcripts consist of four tiers:

Comments	Transcriber's notes (seldom used)
Uncertain	
Non-speech sounds	e.g., inbreath, outbreath, voice onset, sniff, giggle, door slam, silence, channel noise, cough, whistle, yawn, room noise, lip smack, click, loud laughter, blowing noise.
Orthography	

- Segmentation at the TCU, word and phone levels.
- Forced alignment used to generate segmentation at word and phone levels.
- For BE, forced alignment used acoustic models and pronunciation dictionary trained on the AMI corpus [1].
- The 20ms error for BE (proportion of boundaries placed more than 20ms from ground truth boundary) was 35%.
- No resources available for BSC, so used cross-language forced alignment [2].
- Alignment used recognisers trained on Czech, Russian, Hungarian or American English (AE); BSC phones automatically mapped to the closest phone of the target language.

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

- Best alignment with Russian recogniser, significantly better than Hungarian and AE (Wilcoxon signed rank test, $p<0.01$).

Corpus Availability

- The corpus will be released from July 2012 under the Creative Commons Licence.
- Accessible via a web-based search engine which allows download of audio and video segments.
- Search for participants, annotated items, regions of overlapping speech etc.

Conclusions and Future Work

- Novel audio-visual corpus of spontaneous multi-party conversations in BSC and BE.
- Cross-language techniques are viable for segmentation of BSC, for which few resources exist.
- Overlapping talk is common in the corpus.
- Current work is investigating how prosodic resources are used to signal competitive vs. noncompetitive overlaps in the two languages [3].

Acknowledgments

Arts & Humanities Research Council

Supported by AHRC. Thanks to Matt Gibson and Thomas Hain for assistance with forced alignment.

References

[1] 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, pp. 357-360.

[2] 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.

[3] Kurtic, E., Wells, B. & Brown, G, J. (2012) Competing for the turn in conversation: A cross-linguistic comparison. BAAP 2012 Colloquium, University of Leeds, 26th-28th March.

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

[5] 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.