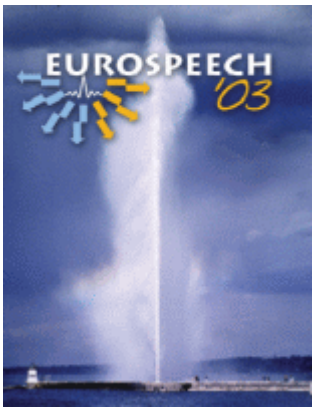
ISCA Archive
http://www.isca-speech.org/archive

EUROSPEECH 2003 - INTERSPEECH 2003
8th European Conference on Speech Communication and Technology

Geneva, Switzerland
September 1-4, 2003



Creating Corpora for Speech-to-Speech Translation

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, Seiichi Yamamoto

ATR-SLT, Japan

This paper presents three approaches to creating corpora that we are working on for speech-to-speech translation in the travel conversation task. The first approach is to collect sentences that bilingual travel experts consider useful for people going-to/coming-from another country. The resulting English-Japanese aligned corpora are collectively called the basic travel expression corpus (BTEC), which is now being translated into several other languages. The second approach tries to expand this corpus by generating many "synonymous" expressions for each sentence. Although we can create large corpora by the above two approaches relatively cheaply, they may be different from utterances in actual conversation. Thus, as the third approach, we are collecting dialogue corpora by letting two people talk, each in his/her native language, through a speech-to-speech translation system. To concentrate on translation modules, we have replaced speech recognition modules with human typists. We will report some of the characteristics of these corpora as well.

Full Paper

Bibliographic reference. Kikui, Genichiro / Sumita, Eiichiro / Takezawa, Toshiyuki / Yamamoto, Seiichi (2003): "Creating corpora for speech-to-speech translation", In *EUROSPEECH-2003*, 381-384.