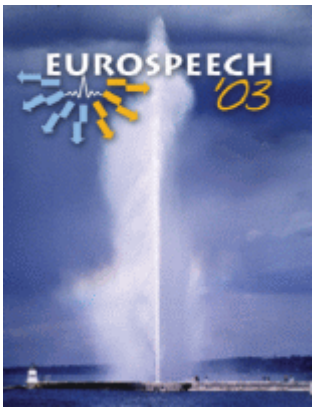
ISCA Archive
http://www.isca-speech.org/archive

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

Geneva, Switzerland
September 1-4, 2003



Lexica and Corpora for Speech-to-Speech Translation: A Trilingual Approach

David Conejero, Jesus Gimenez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Nuria Castell, Asunción Moreno

Universitat Politecnica de Catalunya, Spain

Creation of lexica and corpora for Catalan, Spanish and US-English is described. A lexicon is being created for speech recognition and synthesis including relevant information. The lexicon contains 50K common words selected to achieve a wide coverage on the chosen domains, and 50K additional entries including special application words, and proper nouns. Furthermore, a large trilingual spontaneous speech corpus has been created. These corpora, together with other available US-English data, have been translated into their counterpart languages. This is being used to investigate the language resources requirements for statistical machine translation. Se describe la creacion de lexicos y corpus para el catalan, castellano e ingles hablado en Estados Unidos. Un lexico conteniendo informacion relevante para el reconocimiento y sintesis del habla esta siendo creado. El lexico contiene 50.000 palabras comunes seleccionadas con el fin de lograr una amplia cobertura de los dominios escogidos, y 50.000 entradas adicionales que incluyen vocabulario especifico, y nombres propios. Ademas, se han creado corpus orales para el catalan y el castellano. Estos corpus, junto con otros datos disponibles sobre ingles hablado en Estados Unidos, han sido traducidos a las otras dos lenguas con el proposito de generar un gran corpus trilingue. Este esta siendo utilizado para investigar los requisitos de los recursos linguisticos para la traduccion automatica estadistica.

Full Paper

Bibliographic reference. Conejero, David / Gimenez, Jesus / Arranz, Victoria / Bonafonte, Antonio / Pascual, Neus / Castell, Nuria / Moreno, Asunción (2003): "Lexica and corpora for speech-to-speech translation: a trilingual approach", In *EUROSPEECH-2003*, 1593-1596.