The AIM Corpus: Some notes on the electronic collection of interpreter-mediated interaction

While corpora of spoken language have increased enormously in the last years and their availability is now large, methods of analyzing such corpora electronically do not seem to have followed accordingly.

Corpora have traditionally grown in the realm of phraseology (Sinclair 1991), and, within this realm, attempts to apply concordance analysis to speech are quite an exception. Aston (2015) for instance is one of the few studies focusing on recurrent speech prosody of phraseological items. The corpus discussed by Aston is a small set of TED talks, which he has experimented in teaching (and learning) conference interpreting.

In the area of Conversation Analysis, talk has been collected since the first works of Harvey Sacks in the 1970s, but reasoning about how to implement such collections of talk for electronic analysis is very recent and experimental (Stivers 2015; Steensig and Heinemann 2015).

For the last 15 years, we have collected spoken data for research in a relatively novel area of studies that has become known as Dialogue Interpreting (DI, Mason 1999). Dialogue interpreting consists in interactions involving speakers of different languages and bilingual participants who translate to allow their interlocutors to communicate with each other. Our corpus of data has been collected specifically in healthcare settings located in highly industrialised migration areas in North Italy. The conversations in our corpus normally involve a health worker (doctor, nurse o midwife), a migrant patient and a so called language-cultural mediator, a bilingual speaker with both an interpreting and a migration experience.

The following are, in short, the data we have gathered at the moment. The corpus is called AIM (Analysis of Interaction and Mediation) after the name of the national group of researchers who have collaborated to its implementation and analysis (http://www.aim.unimore.it/site/home.html).

The AIM corpus (2004-2018)

Language couples	No. encounters	Recording time	No. mediators
Italian-English	262	2587’	4
Italian-Arabic	163	2265’	5
Italian-Chinese	81	1195’	2
Italian-French	22	266’	5
Tot.	**548**	**6977’ (over 100 h)**	**17**

The corpus is a large one in its category and for some years now we have tried to reorganize it in a way as to make it more easily available and searchable (see Niemants forthcoming). We have used the annotation tools EXMARaLDA and ELAN to start recreating our transcripts as audio annotations.

In this paper, we report on this experience by showing our preliminary analysis of a sub-set of 65 encounters, whose transcript is now aligned with audio, for a total amount of 1.124 minutes (almost 19

hours) interaction. Our work has gone in two directions: 1. Exploring which lexical items may constitute interesting starting points to retrieve relevant structures of interaction; 2. Finding ways of extracting audio-aligned transcripts to use in e.g. specific interpreting training or learning activities. In our presentation, we will show and discuss one example of each type of search.

The problems we are coping with, and would like to discuss in our presentation, have to do with a potential change of perspective following the new possibilities we have to search the corpus. While interaction analysis is based on the observation and qualitative analysis of talk sequences, finding "interesting" sequences by corpus search (rather than human eye) opens a series of questions involving data annotation and encoding.

References

Aston, G. 2015. Learning phraseology from speech corpora. In A. Lenko-Szimanska and A. Boulton (eds.) *Multiple affordances of language corpora for data-driven learning*. Amsterdam: Benjamins. Pp. 63-84.

Mason, I. (ed.) 1999. *Dialogue Interpreting*. *The Translator*, 5/2. Special Issue.

Niemants, N. frth. Des enregistrements aux corpus :transcription et extraction de données en milieu médical. *Meta*, 63(3). Special Issue.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Steensig, J. and T. Heinemann 2015. Opening up codings? *Research on Language and Social Interaction* 48 (1): 20-25.

Stivers, T. 2015. Coding social interaction: A heretical approach in Conversation Analysis? *Research on Language and Social Interaction* 48 (1): 1-19.