CORPUS LINGUISTICS METHODS IN INTERPRETING RESEARCH: A CASE STUDY

Šárka Timarová
Charles University, Prague

1. Introduction

Among the central issues in interpreting research methodology is the question of how to approach and analyse experimentally collected data. Computer-aided analysis (cf. Pöchhacker 2004: 199) and corpus-linguistic methods in particular are one possible path (Pöchhacker 2004: 202). The use of corpus managers for analysis of large data files has been proposed more than once in translation studies by Baker who also published several empirical studies with examples of such analyses (e.g. Baker 1993, 1995, 2000). A similar proposal for interpreting studies was made in Shlesinger (1998). In this paper, I would like to describe some contributions and implications of the corpus-linguistic methods for interpreting research, and show two detailed step-by-step analyses to encourage more ideas.

2. Corpus Managers

The first and most obvious advantage of corpus managers (CM), the basic software tool, is their speed and capacity to process large amounts of data. Many previously laborious steps in data analysis can be done as, literally, one-click operations on a large number of data files. CMs also have many in-built functions. Some of those that only require pressing one button include list generation of all words found in the files, in alphabetical or frequency order, basic statistics on the total number of words (*tokens*), number of different words (*types*), number of sentences, average number of sentences per text, average number of words per sentence, number of sentences with 3, 4, 5 ... words. When searching for a particular item (a word, phrase ...), functions such as concordance (displays the item in a context of e.g. 5 preceding and 5 following words) or plotter (shows the distribution of the item throughout the text) are of great assistance. All statistics and searches can be done on a very large amount of individual files at the same time (i.e. all participant outputs). This list, already quite long, still does not cover all basic functions.

Perhaps the major challenge for the use of CMs in interpreting research is the need for availability of the data in an electronic format. This requires that the researcher still undertake a rather laborious transcription of the audio

recordings. At the moment, there are no reliable speech-to-text converting tools for many languages. Also the transcription requires that the researcher stop and think beforehand what exactly she wishes to investigate. Common CMs were primarily developed for processing written texts. This implies that they are not able to "read" the text in other than orthographic form. The transcript of audio output cannot therefore include any extralinguistic features (marks), such as intonation rise or hesitation within a word (e.g. *presi↑dent* would not be recognised as *president*), and a sentence would not be recognised if it does not start with a capital letter and end with full stop. Similarly, any unfinished words will not be recognised as such, but rather as words in their own right (delimited by a space on each end of the letter string, e.g. *I would like to co*[me] *go home*., where the unfinished *co* for *come* would be recognised as a word per se).

CMs can indeed be of great help for quantitative analysis, but one must bear in mind that they are only tools. The possession of an oven and a cookery book does not mean one has a meal, and having the most advanced text editor still means one has to write all papers oneself. Similarly, even with a CM, the researcher must have a very clear idea of what she wants to look for and how to look for it. In the following section, I will describe two analyses with emphasis on all major decisions that had to be made throughout the process to arrive at the desired result.

3. Sample Analyses

Participants

There were 18 participants: interpreting students who had completed their interpreting training and graduates with a maximum of 3 years of professional experience.

Materials

Two genuine recordings of conference speeches in English were used as source texts. Interpreting was recorded on common audio cassettes. A standard MS Office package was used for transcriptions (MS Word) and partial data analysis (MS Excel). WordSmith Tools[1], a corpus manager, was used for data analysis.

1 The corpus manager employed in these analyses was WordSmith Tools published by Oxford University Press. Very helpful tutorials and support materials are available at the author's web pages. Mike Scott's webpage can be found at http://www.liv.ac.uk/~ms2928/wordsmith/screenshots/index.htm.

Procedure
Each participant interpreted two source texts: one consecutively and one simultaneously, from English into Czech (C to A). The output was 36 recordings divided into four groups according to text and mode (text 1 consecutively, text 1 simultaneously, text 2 consecutively, text 2 simultaneously).

3.1. Analysis 1: Text Length

Rationale
As a first step in analysing differences between CI and SI, I decided to measure the length. In interpreting research, length is measured either in terms of words or syllables (word count is more frequent, but some authors have serious reservations, cf. Čeňková, 1988:101-102). As English words are generally shorter than Czech words, and as there are e.g. no articles in Czech, it seemed that a mere comparison of the ST and TT number of words would not be informative. Therefore, I decided to take both counts, words and syllables.

Procedure
The first decision had to be made at the stage of transcription. As I decided early on to use a CM for analysis, it was obvious I would transcribe the texts orthographically. For purposes of measuring the length of the output, I decided to include in the transcription everything the interpreters said, including unfinished words. For purposes of the syllable count, I also had to transcribe some abbreviations (such as *USA*) as pronounced so they might be recognised as three syllables (*u es a*). Transcribed TTs (30,000 words, over 60 printed pages!) were then uploaded into the CM. One click produced an overview of the number of words for each TT and a total for a group of TTs (grouped according to text and mode). This operation took about 10 seconds. Counting the syllables was slightly more difficult, as the CMs are not able to recognise syllables. The decision that was made[2] is a good example of how to come up with a procedure which the tool is suited for. In Czech, all syllables are centered around a vowel, with only two relatively infrequent exceptions of diphthongs. Hence, I asked the CM to find all instances of A, E, I, O, U, etc., regardless of what came before or after them (whether they were at the beginning of a word, at the end of it, preceded/followed by other letters). The total number of instances found was the desired number of syllables (I still had to discount the diphthongs, using the same method, this time looking for AU and OU combinations). The most difficult part was coming up with the procedure: the search and count itself was

2 For this idea I am indebted to Mirek Pošta, a colleague and a corpus-linguistics enthusiast.

again a matter of several seconds. Of course, this method will not work for every language. I am just trying to illustrate how the functions of a CM can be used, and how to adapt a research question into a workable procedure.

3.2. Analysis 2: Lexical Density

Rationale

Lexical density is one of the key quantitative corpus parameters (Stubbs, 2002:39). The parameter is based on the fact that languages are composed of content words which are the primary carriers of meaning (nouns, adjectives, verbs, etc.) and function words (auxiliary verbs, pronouns, conjunctions, etc.). Lexical density is calculated as a ratio of the number of content words to the total number of words in a text and is expressed as a percentage, or

lexical density = 100 x number of content words/ total number of words

Lexical density is known to be higher in written texts than in spoken texts. Within the domain of spoken text, Stubbs (1996) found differences in lexical density between texts delivered in an environment with or without a direct contact with the listener. Genres where there is no feedback from the audience, such as answering machine messages or radio commentaries have a higher lexical density than genres where there is such feedback, such as public speeches or radio discussions (Stubbs, 1996:74). This raises an interesting question as to whether there would be a difference in lexical density between consecutive interpreting and simultaneous interpreting output: the consecutive interpreting environment allows for feedback and contact between the interpreter and her audience, while simultaneous interpreting does not. The following analysis describes a procedure for answering this question using a CM.

Procedure

The same small corpus of 36 samples was used. First of all, the transcriptions needed to be adjusted from the previous analysis (*u es a* back to *USA* to be counted as one word, etc.). The 36 files were uploaded to the CM and using the word list function a list of all different words (*types*) was obtained. The word list function produces a list of all words found at least once in at least one file. For each file, it will show how many times a given word appears in the file. The samples had a total of 30,000 words (*tokens* in CM terminology), but because many of them appeared more than once, there were only 3967 different words (*types*). Hence the CM reduced the total number of words the researcher needed to process by a factor of 7.5. The next step was to separate function and content words: I decided to isolate the function words, as their number is much lower than the number of content words. This step had to be done manually by going

through the list of 3967 words. By deleting the "unwanted" content words from the list, the resulting product was a list containing function words amounting to only 447 items. The list was exported into MS Excel and converted to a text file with individual words separated by a comma and a space. The result was a small text file composed solely of function words, which could be included among the "normal" files and serve as a reference file. The text file was uploaded to the CM and a new word list was generated. Clearly, this time the word list contained only function words, as there were no content words in this file.

The 36 tested files were divided into 4 groups according to mode (simultaneous, consecutive) and text (text 1, text 2). The CM generated a word list for each group, and the four word lists together with the function word list were compared: the CM produced a combined overview of all words from the corpus and their frequencies in each of the five word lists. The overview was exported again to MS Excel and ordered according to the function word list. This produces a list where the first 447 lines contained frequencies of function words, and the remaining 3520 lines with content word statistics were deleted. A total number of function words per group was calculated by adding up all frequencies, and slotted into the modified formula for lexical density, where the number of content words was expressed as the total number of words minus number of function words. As a result, four scores of lexical density (one for each group) were obtained.

While the above description may sound somewhat complicated, the actual procedure is rather straightforward. As in analysis 1, the important step is the operationalisation of the research question.

4. Conclusion

The aim of this paper was to provide some practical examples of the use of corpus linguistics methods and its tools in interpreting research, and hopefully to encourage researchers to explore the possibilities corpus managers have to offer for data processing. While primarily designed for quantitative research, corpus managers can help with some aspects of qualitative research as well. It is my belief that corpus linguistics methodology offers valid tools for interpreting research.

References

Baker M. (1993) "Corpus Linguistics and Translation Studies. Implications and Applications", in *Text and Technology*. Ed. by M. Baker, F. Gill

and E. Tognini-Bonelli, Amsterdam-Philadelphia, Benjamins, pp. 233-250.

Baker M. (1995) "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", *Target* 7: 2, pp. 223-243.

Baker M. (2000) "Towards a Methodology for Investigating the Style of a Literary Translator", *Target* 12: 2, pp. 241-366.

Čeňková I. (1988) *Teoretické aspekty simultánního tlumočení* [Theoretical Aspects of Simultaneous Interpreting], Praha, Univerzita Karlova.

Pöchhacker F. (2004) *Introducing Interpreting Studi*es, London and New York, Routledge.

Shlesinger M. (1998) "Corpus-based Interpreting Studies as an Offshot of Corpus-based Translation Studies", *Meta* XLIII, 4, pp. 486-493.

Stubbs M. (1996) *Text and Corpus Analysis*, Oxford, Blackwell.

Stubbs M. (2002) *Words and Phrases. Corpus Studies of Lexical Semantics*, Oxford, Blackwell.