The Ostrava Corpus of Czech and British Radio Discussions as Material for Cross-cultural Analysis of Communication Strategies in the Language of Media

Sirma Wilamová

Abstract

The article points at some problematic issues concerning the acquisition of spoken data from already existing and available Czech and British corpora such as BNC, CNC, DIALOG in relation to the intended cross-cultural analysis of communication strategies in media language. Problems such as representativeness of discourse types and genres, non-randomness, sufficient size, ready and fast availability of information, its topicality and contextualization has finally led to the necessity to create The Ostrava Corpus of Czech and British Radio Debates. The process of creating the parallel corpora in relation to the required parameters of the corpus data is described in the second part of the article.

Keywords: spoken data acquisition, Czech, British radio discussions, cross-cultural analysis, The Ostrava Corpus of Radio Debates 2005-2007

This article is part of the Czech Science Foundation (GA ČR) grant-funded project number 405/07/0176, *Communication and Textual Strategies in Mass Media, Commercial and Academic Discourse (A Contrastive Analysis of English and Czech Discourse)*.

Introduction

Corpus linguistics as a relatively modern discipline running counter to Noam Chomsky's competence/performance approach to language is in line with relatively long-term trends in linguistics. Starting in the late 1960s and early 1970s, corpus linguistics was connected with a pragmatic turn away from a purely linguistic and structuralist focus on language to studying language as social action. This change in approach has logically brought to linguistics a focus on 'real-world' data, and as a consequence also the need for new theoretical venues and approaches. Many of these, such as Conversational Analysis, Critical Discourse Analysis, or Discourse Analysis, observe the communication of people in natural settings and aim at identifying recurring patterns in communication which are interpreted in the co-text and wider situational, social and cultural context through pragmatic analysis.

Problems of data collection

The influx of natural data is also closely connected with computational linguistics, which has considerably influenced contemporary research in many branches of linguistics, using corpus data also as a basic source for modern corpus-based dictionaries and grammars as well as representing a valuable source for interdisciplinary research. In order for corpus data to be valid, it has to have several key attributes such as "[…] typicality, objective nature, non-randomness, sufficient size and ready and fast availability of information" (Daneš 265).

No matter how important corpus data is for linguists today, still there are a number of general problems that have to be faced and which have not been solved yet. Firstly, many corpora are time-limited, which means that they cover data from a certain period of time. Unfortunately they are very rarely modernized or enlarged, hence lose their validity and

utilizability. Secondly, a great problem is the representativeness of the data. This is a crucial issue which shows a certain disbalance and randomness in the structure and the proportion of discourse types in many existing corpora, e.g. BNC, London-Lund, CANCODE, to name just three (Warren 2006). Not only is there a rather striking though understandable prevalence of written language data compared to spoken data, which is obviously caused by numerous difficulties connected with obtaining and processing spoken language, but there is also a disproportion between discourse subtypes (professional, pedagogic, intimate etc.) or genres such as conversation, academic discourse, business discourse, public speeches etc. (Daneš 2003), which can influence the objectivity of the results. A detailed and comprehensive analysis of spoken language corpora is discussed by Warren. The last problem that should be mentioned is very often a lack of contextualization, which is necessary in many linguistic areas but especially in pragmatic research where the context as a dynamic factor is crucial for a correct pragmatic interpretation.

The present state of cross-cultural research

The aim of my research in the team grant project No. 405/07/0176 *Communication and Textual Strategies in Radio, Magazine, Commercial and Academic Texts (a contrastive analysis of English and Czech discourse)* supported by the Czech Science Foundation is to conduct a cross-cultural analysis with the objective of identifying, analyzing and comparing the range of communication strategies used in the contemporary language of Czech and British radio discussions.

Generally speaking, cross-cultural analyses of spoken language in this area are still considerably rare, not to say exceptional, perhaps because there are a number of problems as mentioned above that have to be faced when collecting parallel data in order for the data to be representative as to the discourse types and subtypes under investigation, their broadcast content and the size of the data in order to provide relatively valid conclusions.

A significant work devoted to Czech spoken media language has been published by the renowned Czech researchers Hoffmannová, Čmejrková and Müllerová. The first two of these linguists also participated in the international project on *Czech and Slovak Public Oral Speech in the 1990s*. In their book entitled *Language, Media, Politics* (2003), they and their Slovak colleagues reflect a significantly changing situation in the genre of public oratory, radio and television political and media discourse (interviews, debates, polemics). They offer an interesting analysis of the main trends in the speech situation during the transition period of newly established democratic societies immediately after the fall of the communist regime in 1989.

Also an international team of Czech and Slovak linguists led by M. Ferenčík (University of Prešov) and supported by the VEGA Grant Agency in the period of 2006-2008 (under the Ministry of Education of the Slovak Republic and The Slovak Academy of Sciences) focused on cross-cultural research into politeness in the language of the media.

The problems of suitability of existing Czech and British corpora of spoken language

A major problem when attempting to study authentic spoken language is to acquire suitable (and in the case of a cross-cultural analysis, also parallel) data. With the specific framework of the research in mind, the criteria as to the type of media language –radio discussions, a higher number of participants, a greater variety of topics discussed and varying levels of formality, have been defined.

The very first step was to become acquainted with already existing and available corpora in both languages. These were the 100 million word British National Corpus (BNC), the

Czech National Corpus (CNC) and the DIALOG corpus containing Czech discussions. What turned out to be the main problem and a most serious obstacle revealed only after a thorough search of the existing corpora was the pre-set and specified parameters for the intended analysis.

As for the BNC, only 10% of the texts are made up of spoken material. These are randomly chosen as far as the selected genres are concerned, covering only informal conversations, formal business or government meetings, radio shows and phone-ins. Moreover, unless the corpus is purchased it is not really possible to find out exactly whether the radio programmes available will meet the required criteria, how many will be at disposal and when they were recorded. The BNC material was collected in the years 1991-1994 with a slight (although unfortunately unspecified on the BNC website) revision in 2001 and 2007, which is supposedly connected with its transition to new software tools rather than with the content update. Additionally, considering the character of media discourse, thirteen years or even more is problematic in regard to its topicality.

Another issue is the exact form of the transcripts, which cannot be identified because the examples of longer stretches of texts are not available on the web and cannot be decoded from rather vague information on BNC web pages saying that "The spoken part consists of orthographic transcriptions of unscripted informal conversations (...) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins". (BNC. Retrieved February 2, 2008, from http://www.natcorp.ox.ac.uk/corpus). It is not clear whether the form of the transcripts means purely the orthographically transcribed texts from the recordings, or whether it contains any transcription marks relevant for conversation or discourse analysis such as overlaps, the length of pauses, prosodic phenomena such as intonation etc., because these are more relevant for discourse analysis than e.g. tagging (i.e. formalized marking of particular grammar theory providing morphological, syntactic, lexical, stylistic and other type of information).

The situation with the Czech National Corpus (CNK) is much more positive in many aspects. Spoken data consists of Prague and Brno corpora that have entered the second hundred million of words (Daneš 265) and are being continuously and systematically extended. Authentic recordings of different types of spoken language that differ in their extent, time, content, genre and territorial span are fully available to researchers, which is a great advantage. Apart from the above-mentioned differences, the compilers of the CNK otherwise follow the same criteria, mainly the principles of transcription (Čmejrková, Jílková, Kaderka 244).

Similarly to the BNC, the Czech National Corpus consists of a majority of written material, however the spoken part is continuously being extended and as exact numbers of words are not available, the situation may be different today. The problem, however, is that in the Prague (1988-1996) and the Brno corpora of spoken Czech (1994-1999), the main focus is centred on two discourse types, namely informal conversations between participants who know each other well, and the formal discourse is represented by a fixed and structured question-answer. The newest one million ORAL2006 corpus (2002-2006) also contains an informal conversation between acquaintances or friends (Český národní korpus. Retrieved February 2, 2008 http://ucnk.ff.cuni.cz/), which is the main reason why CNK could not be used for the research.

The only corpus targeting current Czech media discourse has been created at the Institute of the Czech Language, Academy of Sciences of the Czech Republic. The two million word DIALOG corpus contains all genres of media dialogue, focusing mainly on political interviews, debates, polemics, and informal talk shows. It contains valuable authentic material that has been recorded since 1997. Since 2003 it has been revised, and importantly it is

transcribed with a set of conventional transcription marks used for conversation and/or discourse analysis (Korpus DIALOG. Retrieved January 31, 2008 http://www.ujc.cas.cz/oddeleni/index.php?page=DIALOG). Nevertheless, the major drawback for my research is that it covers only television (not radio) discussions and debates.

The Ostrava Corpus of Czech and British Radio Debates recorded in the years 2005-2007

Taking into consideration numerous problems of (un)suitability of spoken data in existing Czech and British corpora in relation to the parameters set for the intended cross-cultural research such as radio discussions from different programmes, a wider range of comparable broadcast content, a higher number of participants and a varying level of formality led me finally to the necessity to build up a specifically designed corpus that would address the above-mentioned criteria.

The Ostrava Corpus of Czech and British radio discussions came into existence in the course of the first year's duration of the grant project, although it is a result of a two-year period of work (2006-2007). It presented a time-consuming and demanding process involving searching numbers of radio programmes on both public service stations, the selection of particular discussions, downloading, for a majority of material writing the orthographic texts from recordings corrected by a native speaker, and transcribed using conventional transcription marks providing special linguistic and non-linguistic information such as identification of the speakers, overlaps, immediate linking of the following utterance, hesitation phenomena, pauses, unfinished words and sentences, repetitions, prosodic features such as intonation, prominence, as well as other relevant comments by the author of the transcript.

The structure of the corpus:

The already existing source material for the research consists of 18 radio debates (lasting 15-45 minutes) in approximately 8 hours of spoken language (230 minutes and 238 minutes for each language) recorded and transcribed. The corpus total is 80 457 words taken from British and Czech radio programmes with comparable content broadcast by public service stations (BBC, Český rozhlas). For practical reasons the corpus uses a transcription convention that is compatible with both the DIALOG corpus and with common conventions used in conversation and discourse analysis.

The criteria for the selection of radio discussions are as follows: (1) a higher number of participants, i.e. interaction between one or two presenters and two or more guests, which ensures a greater variability of relationships; 2) varying content (political, social and cultural topics) in order to give the corpus a wider range of topics). It is to be expected that some topics will bring general consensus among participants, while others will bring disagreement and confrontation, probably resulting in the use of different communication strategies; 3) varying levels of formality determined by the topic and the selection of guests in the studio.

As for the broadcast content, the corpus data are taken from 8 different radio programmes with usually two or more discussions within a single type and different hosts, so that both conventional as well as habitual language behaviour of the different hosts can be observed.

The corpus contains not only discussions about 'serious' topics such as politics, technology, education or science aimed at the whole of society, but there are also radio programmes targeted at specific groups such as women or the disabled discussing their specific problems and issues, which represents a very specific type of programmes with a significant impact on the conversational structure as well as on the discourse strategies used.

The corpus data was collected with the aim of monitoring not only the referential (i.e. informative) but also the affective function, because not only do the media provide and are expected to provide a free flow of news and information, but relatively new trends in media discourse increasingly reflect the efforts to mix the public world of politics, science and education with attributes of private and phatic communication in the otherwise fixed institutional structure of media discourse. The evidence of this trend is apparent in what Fairclough (9) calls 'conversationalization' of media language, which shows that the main emphasis has been partly shifting from information to entertainment.

Finally, in order to provide the opportunity to investigate a wider range of potential communication strategies, the corpus covers programmes where the atmosphere is cooperative and friendly, as well as those programmes where the tone is argumentative or even conflictive and confrontational. This is enabled not only due to the relatively wide range of topics discussed, but also due to the number of participants - ranging from three to six - as well as due to their roles and personal characteristics.

Much has been said and argued about the optimum ways of spoken data acquisition and about their suitability, which is crucial mainly for authentic conversation but can also - to a large extent - be successfully applied to institutional discourse (Roger 1989, Tyler and Cameron 1987, Cheng 2003, Warren 2006). Although I am aware of the fact that the size of the corpus designed for this study cannot be compared to the size or range of discourse and text types of the other corpora discussed before, I believe that its main value lies in the fact that it has been collected for the specific purpose of the research and as such can represent valuable material for a cross-cultural analysis of contemporary language in Czech and British public debates as a specific media genre.

Bibliography

BNC home page. 2 Feb. 2008, <http://www.natcorp.ox.ac.uk/corpus>.

Český národní korpus. 2 Feb. 2008, <http://ucnk.ff.cuni.cz/>.

Čmejrková, S., and J. Hoffmannová, eds. *Jazyk, média a politika.* Praha: Academia, 2003.

Čmejrková, S., L. Jílková, P. Kaderka. "Mluvená čeština v televizních debatách: korpus DIALOG." *Slovo a Slovesnost* 65 (2004).

Daneš, F. "Today's Corpus Linguistics: Some Open Questions." *International Journal of Corpus Linguistics* 7.2 (2003), 265-282.

Fairclough, N. *Media Discourse*. London: Edward Arnold, 1995.

Korpus DIALOG. 31 Jan., 2008, <http://www.ujc.cas.cz/oddeleni/index.php?page=DIALOG>.

Roger D., and P. Bull, eds. *Conversation*. Philadelphia: Multilingual Matters Ltd., 1989.

Taylor, T .J.,and D. Cameron. *Analysing Conversation: Rules and Units in the Structure of Talk*. Oxford: Pergamon Press, 1987.

Address:
University of Ostrava
Faculty of Arts
Dpt. of English and American Studies
Reální 5,
701 03 Ostrava
Czech Republic
Sirma.Wilamova@osu.cz