Spoken Corpora Design Revisited

František Čermák[1]

Abstract

This contribution's aim is to assess most of criteria used in the design of spoken corpora and to arrive at a coherent picture of what is central here and what, being less central, eventually merges with written corpora.

First, (1) different situation of the spoken variety of language vis-à-vis the written language is noted, this being due to a variety of factors, including prescriptivism, state of literacy, etc.

In consequence of this and because of (2) a number of goals pursued, a variety of spoken corpora is to be found. These goals have influence on the type of the spoken corpus (to be) created. The estimated ratio of spoken-written language being 90 : 10%, which may not be far from being true, is a compelling reason for spoken corpora formation.

There is always some (3) difference between the written and spoken language. In reality this varies and depends on the language in question, social tradition, etc. It is argued that where these overlap most, the overall spoken character of a corpus is weak. On a practical side, availability of some texts and a necessity to record other influences character of the spoken corpora.

In contrast to a well-known bipolar approach based on the demographic-contextual type of texts, at least 5 distinct aspects (4) are suggested as basic for the strategy and choice of spoken texts. These include (a) demographic aspects proper (age, etc.), (b) contextual-situational (proximity, etc.), (c) geographic, and (d) discoursal aspects (one-to-one, etc.), as well as (e) orientation (un/scriptedness, etc.).

Using as a basis criteria (b), (d) and (e), a list of nine protypical features (5) of oral corpus is suggested. These may be viewed as a core. Sharply contrasting to these prototypical spoken features, a set of nine features characterizing the written texts is offered (MINUS, below), too. These are:

	A PLUS	B MINUS
1	spontaneous	prepared
2	spoken	read
3	interactive	unidirectional, non-dialog
4	partners present	partners distant
5	dialog	monologue
6	private, non-public	public
7	proximity of partners	no proximity
8	equality of partners	non-equality
9	no awareness of the purpose of recording	awareness

[1] Institute of Czech National Corpus, Charles University, Prague
e-mail: frantisek.cermak@ff.cuni.cz

It is suggested that a prototypical text should have all of the A PLUS features. Growing absence of some of these is a basis for a graded classification; if none is present, these having been replaced by B MINUS features, one is no longer dealing with a spoken text. Along this scale from a full representation of A features to zero a number of transitions to written texts are to be found.

Finally, some open outstanding questions (6) are discussed.