> p. 85 miss what is unexpected about the language under study. In so doing, ↳ we risk depriving the speakers' descendants of what is special about their heritage, and we lose opportunities to expand our own theoretical horizons. (2001: 53)

3.5.3 Text collections

Since 'linguistic elicitation is artificial even under the best circumstances' (Samarin 1967: 59), any grammatical analysis and description should in the first place be based on a good text collection, while the elicitation of data should only be used at the very beginning of the project or as a means of filling gaps in the data as they usually occur in inflectional paradigms (Dixon 2010: 321–2; see §3.5.2.4 above). Correspondingly, the examples illustrating grammatical categories and constructions in a grammar should as much as possible be quotations of naturally occurring utterances (Bright 2007: 16; Mithun 2007a: 59–60, 62–4; Weber 2007: 200). If the corpus does not provide a simple example to illustrate the grammatical phenomenon in question, the grammarian can resort to an elicited example and supplement it by a quotation from the text corpus.

In view of this central role of text collections, the fieldwork manuals contain surprisingly little information on what constitutes a good text collection and how it can be gathered, why it should cover various genres, and what kind of linguistic data can be found in texts of different genres. The fullest accounts of what constitutes a good corpus and what kind of texts it might contain are given by Samarin (1967: 55–68) and Rivierre (1992: 56–63). In the following I will only discuss the content of conventional text collections and how it relates to the morphological and syntactic analysis of the target language; for the technical aspects of recording see Margetts and Margetts (Chapter 1 above), Austin 2006, Schultze-Berndt 2006, and Seifart 2006; for experimental and stimuli-based techniques of recording connected discourse see Majid (Chapter 2 above).

3.5.3.1 Features of a good corpus

Samarin discusses 'six of the outstanding features of a good body of data' (Samarin 1967: 55–68). A good corpus is:

1. 'dialectally uniform';
2. 'natural', i.e. produced and accepted by native speakers as 'appropriate under a given set of circumstances';
3. 'varied', i.e. it would ideally cover all varieties of language that can be attributed to (a) the age, (b) sex, and (c) social class or occupation of the speaker, (d) the emotion at the time of speaking, (e) the speed of utterance, and (f) the topic, (g) type, and (h) style of discourse;
4. 'complete' in that 'all the closed classes of linguistic elements are fully accounted for';
5. p. 86 ↳ 'repetitious' in order to facilitate the identification of the distribution and function of particular grammatical elements;
6. 'interesting', i.e. containing authentic genres and telling something about the culture of the speech community.

The native speakers' use of particular grammatical categories and constructions is determined not only by the structural properties of the language but also by the nature of the particular communicative event, because all languages provide for alternative ways of expression and rules for their contextually appropriate selection. For an introduction to the ethnography of speaking, see Hill (2006), Franchetto (2006), Trudgill