elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 3*. Editor: Peter K. Austin

Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke

FREDERIKE LÜPKE

Cite this article: Frederike Lüpke (2005). Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke. In Peter K. Austin (ed.) *Language Documentation and Description, vol 3*. London: SOAS. pp. 75-105

Link to this article: http://www.elpublishing.org/PID/037

This electronic version first published: July 2014

Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke

Friederike Lüpke

1. Relevance of field-based corpora*

Until recently, field-based corpora constituted a marginal output of field linguistic research: the main concern of the discipline was to create linguistic descriptions in the form of reference grammars, dictionaries, and scholarly papers. Primary data were occasionally published as text collections or as back matter of descriptive works, but generally were only a data mine for the illustrative examples in publications that focused on analysis rather than on providing primary data.

Field-based corpora are suddenly gaining a new importance due to the emergence of the field of documentary linguistics, going hand in hand with, and facilitated by, new technological possibilities in the domains of data storage and access. Rather than being the facultative appendix to any form of linguistic description, these corpora are more and more being seen as outcomes of linguistic research in their own right. ELDP and DOBES, the two largest funding bodies for the documentation of endangered languages, encourage the publication of (mainly electronic) corpora, which, if fulfilling certain conditions, have an enormous potential as an invaluable record of language and culture and as a database for the speech communities as well as for present and future multidisciplinary research.

While the creation and use of these corpora opens up radically new possibilities, it also presents new challenges for field linguists and any potential users. It is the aim of this paper to outline a number of areas in which field-based corpora can make contributions to different linguistic and neighbouring disciplines as well as to language maintenance and revitalisation efforts. The paper further discusses some of the methodological, theoretical, ethical and practical difficulties of which I am aware in order to engage in a dialogue with fieldworkers and specialists from other disciplines on how these issues can be addressed in language documentation.

The data used in this paper to illustrate the major lines of argumentation are the result of eleven months of fieldwork on the Central Mande language Jalonke of Guinea. For more information on the language, its speakers and its structure, the reader is invited to consult Lüpke (2005, 2006). The following section introduces the Jalonke

* Versions and parts of this paper were presented at the BAAL colloquium 'Joint efforts, shared benefits – advances in methodology, good practice and theory building through and for language documentation' in September 2004, at the ELAP workshop 'Multidisciplinary approaches to language documentation and description' at SOAS in November 2004, and at the ELAP workshop 'The what, how and why of data collection in the field' in June 2005. I thank the audiences of these workshops for their feedback. I am indebted to the speech community of Jalonke in Saare Kindia, Guinea, for their hospitality, and to the Max Planck Society for the Advancement of Science, Munich for having funded the field research reported here.

Friederike Lüpke (2005) Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke. In Peter K. Austin (ed.) *Language Documentation and Description Vol 3*, 75-105 London: SOAS

corpus and the types of data it contains in some detail, in order to lay the ground for a discussion of the scope and limits of the usability of corpora assembled by a single researcher in a limited amount of time and based on field research in small speech communities.

2. Description of a field-based corpus

It is obvious that every field-based corpus is unique because of the differences in research interests of the field linguist(s), length of fieldwork, funding, profile of the speech community, availability of consultants, etc. Therefore, the usefulness of such corpora can only be stated in very general terms or based on an individual example, with the consequence that some of the contributions of the specific corpus discussed cannot be extrapolated to any other corpus. Nevertheless, I base the following discussion on my Jalonke corpus in order to benefit from the argumentative force of illustrative examples, without claiming that all field-based corpora should contain the kinds of data present in the Jalonke corpus or that they should aim at answering the research questions underlying the Jalonke corpus. It should be noted that the issue of cross-linguistic comparability of such corpora is an important one that will be addressed in section 4 below.

2.1 Types of data

In the course of fieldwork on Jalonke, a text corpus of approximately 21 hours of recorded speech was collected (see Table 1 for details), consisting entirely of 'observed' and 'staged communicative events' (Himmelmann (1998)). Observed communicative events comprise all texts for which linguists' influence on content and linguistic structure of the utterance is limited to their presence as observers. Staged communicative events are, according to Himmelmann (1998: 185), "enacted for the purpose of recording". Within this category, staged events where speakers are prompted linguistically to talk about a certain topic can be differentiated from events prompted by "props" (Himmelmann (1998: 185)) or stimuli.

Table 1: Structure of the Jalonke text corpus

Communicative event	**Genre**		**Recording time in minutes**
Observed	Narrative	Historical	118
		Personal	226
		Story	127
	Conversation		259
	Other (speeches, songs, proverbs, procedural texts, etc.)		318 (+written collection of pro-verbs and letters)
Staged	Action descriptions		235
Total recording time			1283

A considerable amount of data used in this study was gained through the use of nonverbal stimuli, mainly devised by the members of the Language & Cognition Group at the Max Planck Institute for Psycholinguistics in Nijmegen (see Table 2 for details).

Table 2: Stimuli of the Language & Cognition group used

Stimulus	**Nature**	**Task**
ECOM-Clips (Bohnemeyer and Caelen, 1999)	video animation	preferred and possible description of motion events
The Topological Relations Picture Book (Bowerman and Pederson, 1993)	line drawings	description of topological relations
Picture Series for Positional Verbs (Ameka et al., 1999)	photos	description of topological relations
Enter/Exit Animation (Kita, 1995)	video (animation)	description of enter/exit scenes
New Tomatoman (Allen et al., ms.)	video (animation)	description of motion events
Caused Positions (Hellwig and Lüpke, 2001)	video (acted)	description of caused and spontaneous putting-events
Event Triads (Bohnemeyer et al., 2001)	video (animation)	non-linguistic similarity judgments of motion events and linguistic description of motion events
Demonstrative Questionnaire (Wilkins, 1999)	abstract scenes serving as the basis for re-enactment	reenactment accompanied by speech
Cut & Break Clips (Bohnemeyer et al., 2001)	video (acted)	description of cutting- and breaking-events
Table Top Route description (Wilkins and Danzinger, 1999)	object manipulation	director-matcher task for retracing routes
TEMPEST films (Bohnemeyer, 1998)	video (acted)	description of the temporal order of events

These stimuli are aimed at the collection of cross-linguistically comparable data on a range of research topics, from spatial organisation to the linguistic encoding of motion

events. Most of them are designed to trigger linguistic descriptions of pictures and videos. Besides the MPI-tools, other visual stimuli were used. They are listed in Table 3.

Table 3: Other stimuli used

Name	Nature	Task
Frog Story (Slobin, 1993)	picture book	re-telling the story
The Pear Film (Chafe, 1980)	video (acted)	re-telling the story
The Chicken Film (Givón, 1990)	video (acted)	re-telling the story
Maus films (cartoons from "Die Sendung mit der Maus")	video (animation)	re-telling the scenes
Canary Row scenes (sequences from a Tweedy Bird cartoon)	video (animation)	re-telling the scenes
Quick & Flupke (Tintin)	picture stories	re-telling the scenes

Video-recordings of diverse activities recorded at the field site were also deployed with the intention of collecting online-descriptions covering a wide range of culture-specific contexts. All of the resulting descriptions labelled 'action descriptions' and described in detail in 2.3 are present in the corpus. Staged communicative events also comprise other texts where I had some influence on the content, although not on its linguistic structure, by prompting speakers to talk about certain topics or asking them questions. The text collection resulting from observed and staged communicative events is henceforth referred to as the 'corpus'[1].

A separate database used for statistical analyses consists of ca. 5,000 clauses from thirty different speakers. This database, referred to as the 'sample' throughout the paper, forms a subset of the corpus. The structure of the sample is described in detail in the following section.

2.2 The quantitative sample

A subset of the corpus introduced in 2.1 above was compiled as a sample underlying a quantitative study. This discourse study is based on 7,063 intonation units, featuring 30 different speakers. With the 79 intonation units uttered by me subtracted, the sample contains 6,984 intonation units. These units were divided into clauses, that is, predications with either a verb or a predicate nominal or adposition as its head. Fragments not containing a predicate, false starts, incomprehensible turns, and utterances in French, Fula or Soso, or containing a main predicate in one of these three

[1] Another source of data used for analysis consists of elicitations in the form of translation equivalents, acceptability judgements and contextualising elicitations. These data are not included in the corpus but informed and refined the analysis of Jalonke and helped to establish a lexical database of ca. 2,000 items

languages were excluded. The remaining cases add up to 5,806 predications. The speakers, genres and texts of the sample were selected according to the following criteria:

- The inclusion of a sufficient number of speakers of different ages and genders with a substantial amount of utterances.
- The presence of some speakers with more than one text and ideally also with texts belonging to different genres.
- A representative selection of different genres and texts.

In view of the methodological and practical issues influencing the design of a field-based sample, these points are addressed in some more detail here:

1. In order to reach a representative amount of data per speaker, and to be able to compare across speakers and genres, approximately 100 predications, verbal and nonverbal, were set as a goal per speaker, with minimally 20 speakers in as many genres as possible. Reaching this aim for 20 different speakers meant that for some speakers, the number of included predications had to be much higher, since many of the texts are interactive. Furthermore, some speakers participate in a great number of texts, while others feature only in one instance. Building a collection as representative as possible for 20 speakers from a field-based corpus also meant that the total number of speakers had to be higher than 20, since some speakers only appear with a few utterances in the sample but participate in the selected texts.

2. A further attempt was made to represent different genres and/or texts from some speakers in the sample, so that consistency or deviations not only between speakers, but also within speakers across genres and texts could be assessed. Given that many field-based corpora are skewed through the dominance of a limited number of main consultants, and that by far the most 'popular' genre contained in corpora tends to be story telling, the attempt to include several speakers with different genres had to be limited to a small number; when it came to represent one and the same speaker with different texts from the same genre, this goal turned out to be impossible to reach for most genres.

3. The sample was designed to contain the widest possible variation of genres, in order to be balanced with respect to differences in information structure, packaging, planning, etc., known to vary with genre. Again, it has to be conceded, however, that many genres could only be represented with one speaker and often only with one text, so that some of the generalisations reached on the basis of these data must remain preliminary and do not lend themselves to statistical tests. Within genres, attention was paid to varying the topics of texts as much as possible with the intention of covering the widest possible range of real-world situations and events, and hence of linguistic descriptions of these situations and events.

Finally, it is important to stress that none of the staged communicative events resulting from the stimuli listed in Table 2 and 3 above feature in the sample. The reason for excluding stimulus-based texts from the sample is that although they consist of communicative events on whose linguistic structure I did not have direct influence, they do not constitute examples of entirely natural language use. Many of the stimuli investigating specific linguistic domains yield repetitive use of the same constructions, for instance, and none of them triggers descriptions that are common cultural practices of Jalonke speakers. Nevertheless, some of the action descriptions briefly introduced above are contained in the sample. The reason for including this genre in the sample is that, although a novel genre invented for the purpose of linguistic investigation, they are very close to monological narratives and procedural texts in structure. In addition, the texts describe scenes that are part of everyday life in the Jalonke speech community and do not confront the consultants with unfamiliar images or tasks. Therefore, the resulting descriptions are at least very close to communicative events that are part of everyday language use.

The structure of the sample resulting from these considerations is given in Table 4:[1]

Table 4: The structure of the Jalonke sample

Genre	**Text No.**	**Topic**	**Speakers**	**Number of predications**
Conversation	1	Varied	AB	105
			Alpha	253
			A	59
			Abdou	88
			Dian	89
			Assi	4
			Bouba	3
			MCR	3
			Total text	**604***
	18	Women comment on a crocheting lesson given by one of the participants.	ABB	6
			AK	6
			A	99
			anonymous 2	1
			TH	36
			AO	28
			Mai	84
			K	1
			Total text	**261**

[1] More information on the sample structure, including the consultants' age and sex, can be found in Lüpke (2005).

Genre	Text No.	Topic	Speakers	Number of predications
	19	Death in a nearby village; varied	MD	22
			ABB	68
			AK	194
			HD	6
			HB	115
			EH	21
			anonymous 2	22
			TH	4
			AO	2
			Mai	1
			K	33
			L	24
			Total text	**512**
	Total genre			**1377**
Historical narrative	2	How the Jalonke settled in Saare Kindia	TMS	150
			Total text	**150**
	6	About the earthquake in Guinea in 1984	M.Bala	199
			AB	16
			Total text	**215***
	24	About the medical reasons for the scars on the narrator's temples	EHM	45
			AB	16
			Total text	**61**
	Total genre			**426**
Personal narrative	3	About seasonal work in neighboring countries in the youth of the narrator	AA	223
			Total text	**223***
	7	About a journey the narrator undertook	AB	165
			Dian	10
			Total text	**175**
	15	About a meeting the day before to which the narrator went	MAB	150
			Total text	**150**
	16	About how the holiday the following day is going to be celebrated	AK	93
			MAB	14
			Total text	**107**
	20	About marriage procedures and weddings the narrator experienced in the village	ABB	285
			Alpha	114
			AB	102
			Total text	**501***
	21	About the two main narrators' trip to the next town in order to go to the dentist	MD	182
			HD	122
			AK	15
			MSC	50
			anonymous 1	3
			MCR	2
			Total text	**374**
	22	About the narrator's teacher career and a seminar he attended	SKB	367
			AB	34
			Total text	**401***
	23	About the narrator's pilgrimage to Mecca	EHM	183
			AB	15
			Total text	**198**
	27	About the narrator's forced labour under the French	EH	275
			AB	2
			Total text	**277**
	Total genre			**2406**
Story	4	Story about the orphan and the beast	AK	282
			MD	4

Genre	Text No.	Topic	Speakers	Number of predications
			Total text	**286**
	5	Story about the rabbit and the beans	Agi	153
			Total text	**153**
	8	Why the chimpanzee has no house	AB	49
			Abdou	3
			EH	1
			Total text	**53**
	9	Why the chimpanzee has no tail	AB	33
			Total text	**33**
	10	Why the chimpanzee has a red bottom	AB	58
			Abdou	3
			EH	1
			Total text	**62**
	Total genre			**587**
Letter	11	Varied	AB	89
			Total text	**89**
	12	Varied	AB	87
			Total text	**87**
	Total genre			**176**
Action description	13	Live commentary of a soccer game	Abdou	161
			Total text	**161***
	14	Online description of a video showing village women doing laundry at the river	MAB	152
			Total text	**152**
	17	Online description of a video showing village women working in an onion field	M.Bala	153
			Total text	**153**
	Total genre			**466**
Play	25	About the importance of health care	collective	314
			Total text	**314**
	Total genre			**314**
Speech	26	On the occasion of the inauguration of a health post	Alpha	54
			Total text	**54**
	Total genre			**54**
Grand total				**5806**

2.3 Genres

The identification of genres is known to be problematic (Bakhtin 1999 (1986); Biber 1994; Biber 1995; Biber 1998; Ferguson 1994; Finegan and Biber 1994; Hymes 1972; Hymes 1974). Not only can genres be established through a combination of very different social and/or linguistic features, in different terminological traditions they can also contrast or overlap with other categories of discourse classification like register and style. In order to avoid identifying genres based on linguistic features rather than justifying them independently or using universal labels that lack a language-internal basis, I only distinguish genres in a very broad manner. Although I use existing labels as far as possible to identify them, these labels reflect my own intuitions and the culture-specific perspective of Jalonke, except for clearly 'borrowed' genres. The preliminary inventory of genres, roughly equivalent to categories of 'speech events' in

the sense of Hymes (1972: 56), i.e. "activities or aspects of activities that are directly governed by rules or norms for the use of speech", comprises the following:

Stories are a well-established genre in Jalonke. Both men and women engage in the activity of story telling. Stories are highly monological in nature, but display a ritualised minimal interaction between the narrator and the listeners well known in Mande culture in general: A story starts with the following formulaic expressions, sometimes preceded by the exclamation `Taali!` 'A story!' or `Kiini-na` 'The story!' of the narrator:

Narrator:

N	*ma*	*sɛn!*
1SG	?	?

'Here's a story!'

Interlocutor(s):

A	*xa*	*ɲaxun*
3SG	SUBJ	be sweet

'May it be sweet!'

Narrator:

Sube-dii	*a*	*i.*
meat-DIM	3SG	at

'There is a little meat in it.'

Interlocutor(s):

Jaba-dii	*a*	*i.*
Onion-DIM	3SG	at

'There is a little onion in it.'

This exchange is optionally followed by more parallelisms citing additional 'ingredients' to the story. After these preliminaries, the narrator starts the story with the following sentence:

Narrator:

Ndan a ra.

Somebody 3SG with

'There is somebody.'

The narrator then goes on to introduce the main participant of the story. From this point, the only interactive features are occasional questions by the narrator about whether the interlocutors have understood, and, at regular intervals, acknowledgement by the interlocutors that they are following by uttering *naamu*. When the narrator reaches the end of the story, (s)he announces:

Mɛn-na ɲɛn.

There-DEF end

'That was it.'

Stories are referred to as *kiini* or *taali*, 'story', and the activity of story telling is designated by *kiini madɔxɔ* 'story DISTR-sit (down)'.

Historical narratives are very similar to stories with respect to the degree of interactivity. They have a less formulaic beginning, but are very likely to start with the narrator giving his name and Islamic title. Historical narratives are almost exclusively told by men who conserve the oral history and genealogies. Women who participate in this genre are old and of a very high social status, and mostly relate issues like marriage customs in the traditional society. The participation of the audience is limited to `naamu` signalling understanding. What sets historical narratives apart from other narratives is that they contain long genealogical 'lists' clarifying the affiliation of the characters and/or their history of migration. Linguistically, these characteristics are reflected in the recursive use of existential and presentative nonverbal predications in the genealogical lists and/or in the preponderance of the motion verbs 'come', 'go', and 'leave' and of toponyms in texts in which migration is talked about. Historical narratives are designated by the Arabic loanword *taarik* and the French loanword *istɔar*, both meaning 'history'.

Personal narratives share the interactive pattern of historical narratives, but cover much more diverse topics. For stereotypical personal experiences like marrying or the pilgrimage to Mecca, narrative 'templates' seem to exist that make them almost

interchangeable even if uttered by different people and interspersed with uniquely personal features.

Public speeches in Jalonke are virtually absent in the present-day, because the language is confined to the private sphere. I recorded only one speech in Jalonke, which was prompted by my presence as the central addressee. Because of the triglossic language situation, on formal occasions either Fula speakers are present and hence Fula is used, or the specific 'modern' context calls for French. Other planned genres like ritual and religious speech are likewise absent for similar reasons.

Conversation comprises chat and discussion. Conversation is the most heterogeneous genre and can be contrasted with the genres introduced above because it is the most unplanned and interactive and the least cohesive genre with respect to topics. It is also the most difficult to define, because of the absence of clear linguistic features determining it, partially due to the heteroclitic character of conversation – it may be blended with short narratives, riddles, proverbs, stories, expositions, etc. Nevertheless it is a special genre for the Jalonke called *summun* 'chat'.

Genres of which examples were collected but excluded from the corpus are **proverbs** and **songs**. I collected only around 30 proverbs and considered this number insufficient for a quantitative study. Songs were difficult to gloss and translate, because they contain archaisms and a great number of unrecognisable words and phrases (see Barwick, this volume). Other genres well known in Mande culture (Bird 1971; Camara 1976; Conrad and Frank, 1995), and mostly performed by a specialist caste, the 'griots' or bards, are epics, hunter songs and songs of praise. These genres, reflecting oral history or citing genealogies, have died out among the Jalonke, since they have lost their griots.

In addition to these, some recent or novel genres figure in the corpus as well:

Plays are likely to be a relatively new genre in Guinean culture that was introduced and spread by development aid institutions in order to popularise issues like health care, farming methods, and the like. A new generation of stand-up comedians like Jean Michel Kankan whose audio recordings are extremely popular have further contributed to its extension. It is not unusual to see mini-dramas performed at public gatherings, mostly addressing educational issues and employing dark humour. While plays are an established genre in the majority languages and in French, a play written and performed in Jalonke was an initiative of the youth club I initiated in the village where I was based. It is the only genre in the corpus that is both completely planned and highly interactive. Its borrowed character is reflected in the French loanword *teatyr-na* 'theater' to designate it.

Letters, at least in Jalonke, are another novel genre. There is a tradition of letter writing in Fula (employing a modified Arabic script) or in Arabic, gradually being joined by French. Since there is no written tradition for Jalonke, only the consultants I trained in a Latin-based orthography are able to write the language. Because one of them left the village to study and wrote me letters in Jalonke, I had the opportunity to include letters

in the sample. It remains unclear how much these letters are influenced by the existing tradition of letter writing in other languages. Letters are designated with the French borrowing *leter*.

Action description is a genre created for linguistic purposes, i.e. staged linguistic events (Himmelmann 1998) that are based on videos filmed in the village. These videos feature everyday activities, like women doing laundry or cooking a meal, men chopping wood or preparing to pray. The linguistic descriptions resulting from these stimuli are different from all the other genres above because they do not correspond to a speech event type of the linguistic community, and hence they are highly artificial. They consist, however, of linguistically non-prompted utterances and thus differ greatly from elicited data. I included action descriptions in the sample practical considerations concerning the size and composition of the corpus. Observed communicative events (Himmelmann 1998) or naturalistic texts are confined to the above mentioned genres and impossible to vary in a controlled manner according to topic and real-world situations covered. The inclusion of action descriptions greatly increased the topics and the number of lexical items present. Action descriptions include expository texts ('How to plant the garden') and plain descriptions of the filmed situations, which is one of their main advantages over linguistically prompting consultants to describe how they perform certain actions. Because the consultants do not have a narrative goal when presented with a video film which they are asked to describe, they do not resort to cultural scripts (such as 'cultivating') but give very fine-grained event descriptions (digging a hole in the ground with a hoe, planting a seed, filling the hole with ground, using hands, etc.) which uncover many lexemes and constructions otherwise rare or absent from the corpus.

3. Contributions of field-based corpora

To compile, annotate and analyse field-based corpus data for which automated tagging, parsing, etc. is not available in most cases requires an enormous investment in time and human resources. More often than not the project has just one researcher assisted by one or two native speaker consultants to carry out the work. It is a pertinent question then to ask what is the value of such a major undertaking. To answer this let us start by looking at how corpus data can complement data from a qualitative inspection of the domain in question in order to address some central research questions relevant to linguistic theory.

3.1 Contributions to theoretical linguistics

The Jalonke corpus was compiled in order to investigate argument structure classes in the language, i.e. classes of verbs that can be distinguished through the number and syntactic and semantic status of the participants they take. I have chosen the potential contribution of corpus data to this area of linguistics to exemplify the usefulness of corpus data for linguistic theorising.

The general research question was motivated by the observation that languages differ considerably with respect to variation in valence, and that these differences can be related to different phenomena, such as argument ellipsis (see Bickel 2003; Fillmore 1986; Huang 1984; Huang 1989; Li 1997; Li and Thompson 1979; Pu 1997), alternations licensed language-individually (Guerssel et al. 1985; Haspelmath 1987; Haspelmath 1993; Levin 1993; Levin and Rappaport Hovav 2003; Mohammed Guerssel 1985; van Hout 1996, 2000), or general absence of regularities in valence as reflected in discourse patterns (Thompson and Hopper 2001). In addition, these phenomena can be analysed in different ways:

- Through lexicalist approaches to argument structure (Alsina 1996; Bresnan, 2001; Butt and Geuder 1998; Comrie and Polinsky 1993; Grimshaw 1990; Hale and Keyser 2002; Levin 1995; Levin and Pinker 1992; Pinker 1989; Rappaport Hovav and Levin 1998; Rappaport Hovav and Levin 2002)
- Constructionalist accounts of verbs and their participants (Bencini and Goldberg 2000; Fillmore and Kay 1993; Fried and Östman 2004; Goldberg 1995; Goldberg 1997; Goldberg 1998; Goldberg 1999; Goldberg 2002; Goldberg 2003; Östman and Fried 2005).

While lexicalist approaches assume that verbs specify information on the number and thematic roles of their participants in the lexicon, proponents of constructionalist accounts attribute this information to constructions and take verbs not to lexically encode information pertaining to their participants.

In addition, some descriptions of Central Mande languages closely related to Jalonke classify all verbs as labile (for Mandinka see Creissels 1983, 1991), or classify verbs based on their minimal syntactic valence (Tröbs 1998). Researchers working on other languages use labels based on argument structure (‘transitive’) reflecting the unmarked valence of these verbs (Dumestre 1994). These classifications are crucially intertwined with the recognition of a passive alternation: throughout Central Mande language there is a construction in which the object participant of a syntactically bivalent verb is mapped to subject and receives an obligatory Theme interpretation. The Agent participant is then not syntactically expressed in some of the languages or optionally expressed as an adjunct in other languages. There is no formal marking to distinguish the construction from an active intransitive clause. Scholars who analyse the monovalent construction as a passive alternation have no difficulties in recognising a class of transitive verbs. Scholars who analyse the two constructions for the verbs in question (and all other verbs with a range of formally zero-marked syntactic options) as valence lability deny argument structure classes based on unmarked valence and alternations. The admissibility of formally unmarked passives has been questioned universally by, for example Haspelmath (1990) and Mel'cuk (1993), so it is an empirical question on what grounds the passive can be identified as an alternation rather than variable valence, with argument drop triggered by information structure.

In short, different hypotheses, some of them arising from contrasting theoretical assumptions on the level of information structure at which the number and status of participants is specified, need to be tested in order to arrive at a valid classification of verbs. More importantly for the point of the present paper, corpus data are needed in

order to complement the argument structure properties of Jalonke as they present themselves based on morphosyntactic criteria.[2] These properties are independent of the number of arguments realised by the tokens of a given verb type in discourse in the sense that they are based on 'possible' rather than 'actual' numbers of arguments. Thus they can be investigated qualitatively through the unweighted investigation of corpus and lexical data. Only through quantitative discourse data, however, can we examine how, lexical argument structure, assumed and in a sense idealised, is aligned with the number of arguments for a given verb in clauses in discourse. The ratio of possible to realised arguments is indispensable information to decide whether argument structure is a lexical property of verbs or a property of the constructions in which verbs appear.

The following tables give a selection of the main results of the quantitative discourse study. Table 5 lists the frequencies for types and tokens of the argument structure classes established on the basis of morphosyntactic criteria, contrasted with the frequency of occurrence of verb types of the different classes in the Jalonke lexicon of 2,000 entries.

Table 5: Distribution of verb tokens and types in the overall sample and of verb types in the Jalonke lexicon over argument structure classes

Argument structure class	**Sample (verb tokens)**		**Sample (verb types)**		**Lexicon (verb types)**	
	Absolute	%	Absolute	%	Absolute	%
Causative/inchoative	573	11.3	17	6.2	23	5.5
Transitive	2574	50.6	150	54.3	223	52.8
Reflexive-only	88	1.7	16	5.8	24	5.7
Intransitive	1849	36.4	93	33.7	152	36
Total	5084	100	276	100	422	100

The distribution of argument structure classes over verb tokens as well as types in the discourse sample roughly mirrors the distribution of argument structure classes over verb types in the Jalonke lexicon. The high number of transitive verb types in the lexicon and of transitive verb tokens in the sample would make Jalonke an extremely transitive language[3] not only in terms of lexical organisation, but also in discourse patterns.

The distribution of non-alternating verb tokens (i.e. those, for which lexical argument structure and valence match) and alternating verb tokens, categorised by the

[2] The formal properties considered in Lüpke (2005) are: the number and syntactic function(s) of the argument(s) with which a verb can or must occur; the alternation(s) in which a verb participates; the valence-changing operation it undergoes; the meaning change entailed by the valence-changing operation; and the nominalisation pattern for the verb. Based on these features, four argument structure classes can be established.

[3] Nichols (1993: 74) regards the figure of 41% transitive verbs in the Russian lexicon as extremely high.

recognised alternations of Jalonke in Table 6 allows an evaluation of the factors conditioning observed deviations from lexical argument structure.

Table 6: Distribution of alternations in the overall sample

Alternation	**Sample (verb tokens)**	
	Absolute	%
No alternation	**4630**	**92.2**
Applicative alternation	11	0.2
Imperative	98	1.7
Unexpressed O alternation	49	0.8
Passive	296	5.1
Total	5084	100

In 92.2% of cases, no alternation occurs, i.e. lexical argument structure as predicted from independently established features corresponds to syntactic realisation of arguments. With respect to markedness criteria, a look not only at the verb tokens, but also the verb types occurring in the different alternations is worthwhile. The passive accounts for the highest number of verb types of all alternations, as expected given its compatibility with all transitive verbs. Still, the passive is not only far less frequent than the active for the verb types that appear in it, it is also rare with eligible verb types: only 81 of the 276 verb types in the sample appear in the passive. Independently of speaker and genre, only 10.8% of all clauses featuring transitive verb tokens were passives; 84.4% were active clauses with two realised arguments (the remainder of the clauses contained transitive verb tokens in the unexpressed object alternation or in the imperative). This distribution of verb types and tokens makes the passive a highly marked alternation according to two of four markedness criteria – 'raw frequency' and 'degree of productivity' (Comrie 1988). 'Discourse distribution', another of Comrie's markedness criteria, will be looked at in 3.2; his fourth criterion, 'formal complexity' is not applicable because there is no formal marking for the passive. Just 38 verb types occur in the imperative, cross-linguistically known to be eligible for non-stative verbs. In contrast, only four verb types occur in the unexpressed object alternation (out of the five verbs participating in that alternation in the lexicon), and only one verb *wale* 'work' occurs in the applicative alternation in the sample. In the lexicon, two verb types are attested in the applicative alternation. Note that 17 of the 23 verb types analyzed as causative/inchoative alternating in the lexicon appear in the sample.

In addition, there are no deviations from lexical argument structure other than those explainable in terms of alternations. No transitive verbs occur in the unexpressed object alternation, other than the five identified by lexical tests as showing that alternation. Intransitive verbs other than *wale* 'work' do not participate in the applicative alternation in the sample. Moreover, omission of subjects and objects does

not occur except where a lexical alternation is possible. Context shows that verb alternations account for non-occurrence of arguments.

By far the most important finding is the startling distribution of non-alternating and alternating verb tokens. The vast majority, 92.2%, appear with the number of arguments predicted by their argument structure, and causative/inchoative alternating verbs are distributed more or less evenly over the two valence classes for which they are eligible. Thus, alternations are confirmed as insignificant, and the argument structure classes are confirmed. Even an account not based on argument structure and alternations from it would have to recognise the overall marginality of deviating instances. In consequence, Jalonke is a language in which lexical argument structure is an extremely good predictor of the number and status of arguments attested for a given verb type in discourse. Cross-linguistic and Mande-specific approaches that deny the existence of lexical argument structure universally are therefore challenged by the facts of Jalonke. The observations of Bickel (2003) and the findings for Jalonke presented here suggest that there may be systematic differences between languages in the degree to which the information on participants is predictable for a given verb and hence likely to be stored in the lexicon vs. unpredictable and hence likely to be contributed by the constructions in which the verbs occur or by principles of information packaging. It follows that a typology of languages is needed in terms or their 'referential density' Bickel (2003) or the ratio between possible and actual number of arguments attested in discourse. To assess the profile of a language in the domain of argument structure and argument realisation, corpus data are indispensable.

The following section will exemplify how the genres identified for Jalonke in 2.3 can be confirmed through structural characteristics attested across languages. For the purpose of the present paper, only two linguistic characteristics will be discussed: the frequency of the passive alternation and the frequency of the imperative.

3.2 Contributions to genre and register studies

This section illustrates how corpus data can be put to use in the identification of linguistic features that enable the differentiation of genres according to criteria such as plannedness, politeness, and formality.

There is a large variation across genres with respect to the frequency of the different alternations in Jalonke. Table 7 below shows how the passive, one of the most productive alternations, is distributed in the different genres. The passive occurs with the lowest percentage of all verb tokens in the play, and with the highest in letters. Most plausibly, this variation is due to plannedness in the sense of Ochs (1979). Relatively unplanned discourse is characterised by less forethought and organisational preparation than relatively planned discourse. Among the linguistic features that are concomitant with plannedness is voice: according to Ochs (1979), the passive voice is more frequent in planned than in unplanned discourse. Since the play, a planned genre, features the least amount of passives, this claim seems to be contradicted by Jalonke at first sight. The contradiction is resolved, however, if we acknowledge the existence of "planned

unplanned discourse" Ochs (1979:77), that is, discourse that has been planned to give the impression of being spontaneous. Since the play 'simulates' natural interactive discourse, it makes sense to classify it as a genre that is intended to come across as unplanned. Under this analysis plannedness may well account for the distribution of passives; passives are infrequent in the more unplanned genres of plays, action descriptions and conversation. They increase in more planned genres, i.e. stories, personal and historical narratives, and are highest in letters.

This pattern moreover confirms the passive as the marked voice for transitive verbs according to Comrie's markedness criterion of 'discourse distribution'. Passives range from a maximum of 8% in letters to a minimum of 1.9% in plays against an average of 5.1%.

Table 7: Distribution of passives over genres and in the overall sample

Genre	**% of occurrence of the passive**
Play	1.9
Action description	2.4
Conversation	2.8
Story	3.2
Speech	5.6
Personal narrative	7.2
Historical narrative	7.3
Letter	8
Overall sample	**5.1**

Another linguistic feature that is distributed in a quite heterogeneous way over genres is the occurrence of imperatives, illustrated in Table 8. Their frequency seems to be correlated with monologic versus dialogic discourse (Himmelmann (1998)). Monologic discourse is unlikely to contain imperatives, with the exception of reported direct speech, so the frequency of imperatives is related to the amount of reported direct speech. In dialogic genres, the frequency of imperatives is related to the degree of formality and politeness. More formal and polite genres are expected to contain no or few imperatives, but rather subjunctives are used to express demands. These expectations are confirmed: in the polite genres of speech and letters, imperatives do not appear, because requests are framed in the subjunctive. In the one monological genre that does not exhibit reported direct speech, action descriptions, imperatives are absent too. The other genres either feature reported direct speech or contain direct commands to the audience – this is marginally the case for historical and personal narratives, and more extensively so for plays, conversations, and stories.

Table 8: Distribution of imperatives over genres and in the overall sample

Genre	**% occurrence of the imperative**
Speech	0
Action description	0
Letter	0
Historical narrative	0.2
Personal narrative	0.5
Play	1.9
Conversation	3.6
Story	5.3
Overall sample	**1.7**

The possibility to measure the frequency distribution of linguistic features in genres established on the basis of 'speech events' of the speech community in question is an invaluable outcome of corpus data coded for the relevant variables. Such information does not only make a contribution to cross-linguistic genre studies but can also inform the creators of language materials aimed at language maintenance and revitalisation to design materials that reflect patterns of actual discourse. This contribution of field-based corpora will be addressed in section 3.4. Now, I turn to a related topic – the potential that the metadata associated with corpus data have for giving important sociolinguistic information.

3.3 Contributions to sociolinguistics and contact linguistics

As extremely widespread in African context (cf. Abdulaziz (2003); Batibo (2005); Makoni et al (2000); Myers-Scotton (1993)), Jalonke is spoken in a complex and triglossic language situation. Along with Jalonke, which is mainly confined to the private sphere, people in the Futa Jalon in Guinea use the regional language Fula to communicate with native speakers of Fula, who are generally not bilingual in Jalonke. Fula tends to be a language of inter-ethnic communication and of writing (in an Arabic-based script). Speakers of Jalonke who have had sufficient access to formal schooling master French, the ex-colonial and now official language of Guinea, to some extent. Most speakers of Jalonke also list the sister language, Soso, which is dominant in and around the capital Conakry, as a language in which they are fluent.

Because bilingualism is the norm for every speaker of Jalonke and multilingualism is also prevalent, Corpus data on how speakers use the different languages they have at their disposal is of great scientific interest. One advantage of the use of corpus data in this domain is that it comes with information about the speakers rather than being disembodied examples in a descriptive grammar. Corpus data, if encoded according to standards recommended by language documentation archives, comes with descriptive and cataloguing metadata (see Nathan and Austin (2004)) that

allow for the retrieval of information on the speakers, their first and other languages, age, sex, and educational background for instance. This metadata enables us to see how language use is correlated with any of these variables, thus yielding valuable sociolinguistic information. It is also relevant for contact linguistics since it permits us to identify those semantic and usage domains in which languages are likely to be used together and thus test the plausibility of scenarios for contact-induced language change. In the following, the potential of corpus data in these areas will be illustrated through the investigation of code-switching/borrowing[4] across the different speech genres, age and sex groups, and different educational backgrounds. We begin with speech genres.

Overall, Jalonke accounts for 98.1% of the 6,984 intonation units, followed by Fula and French with 0.9% each. Borrowings from Soso are statistically negligible. Borrowed or code-switched lexical matter is unevenly distributed across genres: 41.4% occurs in personal narratives, 38.3% in conversation, 12.8% in historical narratives, 4.5% in stories, 2.3% in letters, and 0.8% in action descriptions.

This distribution reflects two different factors: topic and expression of identity. If the topic is associated with aspects of traditional Jalonke culture, borrowings are generally rare, not only because alternatives to borrowings are more available in these contexts, but also because talking about these topics confirms Jalonke identity, which is indirectly flagged through the low frequency of foreign words. Topic and confirmation of identity are most apparent in the genres speech and play, which are virtually free from loanwords or switched speech. Both genres are only represented by one specimen each, which was created in a conscious effort to (re)conquer domains normally occupied by other languages: public speeches are held in Fula in most contexts or in French in a political context, and plays are generally performed in Fula. The two texts for these genres not only have topics associated with Jalonke culture but obviously were created to express Jalonke identity.

If we also look at the internal distribution of languages within the genres, the following picture emerges:

[4] In view of the difficulty if not impossibility of distinguishing code-switching from borrowing and the different and partly conflicting structural properties assigned to these mechanisms (see Thomason (2001) for an in-depth discussion), I do not differentiate between them *a priori*. Rather, I argue that extralinguistic information on the profile of the speakers can help to tease apart code-switching or use of material from several languages from borrowing or conventionalised and well-established material of foreign origin.

Table 9: Languages used in the 6,984 intonation units of the sample according to genre

	Language of intonation unit in %		
Genre	**Jalonke**	**Fula**	**French**
Speech	100		
Play	100		
Action description	99.8		0.2
Story	99.1	0.8	0.2
Letter	98.4	0.5	1.0
Personal narrative	98.2	0.9	0.9
Conversation	97.2	0.7	1.9
Historical narrative	96.6	3.4	0

Fula as a donor language[5] is attested where regional facts are related, headed by historical narratives which mainly describe disputes between the Jalonke and waves of Fula immigrants and invaders. French occurs in association with modernity and national present-day life, such as formal education, national politics, or new technologies.

Jalonke women are more conservative than men in language use: women's code-switching/loanwords are equally split between Fula and French but, men take much more material from French (89.4% of loans in the male corpus). Men's remaining borrowings are 6.5% from Fula and 4.5% from Soso.

A very plausible explanation for the preponderance of French borrowings with men is that they tend to have more access to formal education and hence are more likely to master French. For Guinea, national statistics suggest that this may be valid: 17.1% of girls as against 33.6% of boys are enrolled in school according to the United Nations Statistics Division (2003). However access to formal education and with it mastery of French alone does not account for the use of French loans. Even those Jalonke speakers without French-based schooling use French loans, albeit much less than those who went to state school, as shown in Table 10.

[5] The identification of Fula loanwords is based on the intuition of the speakers as well as my analysis but could not be confirmed from lexicographic information on Fula, since there is no dictionary for the Futa Jalon variety of Fula.

Table 10: Language of borrowings correlated with formal education

	Language of loanwords/borrowings in %	
	French	Fula
Formal education	72.7	54.5
No formal education	27.3	45.5

If speakers who do not master the donor language use material from it a strong case can be made for analysing this material as integrated borrowings that should be considered part of the receiver language lexicon.

In summary, patterns of language use can be identified through corpus data and present important information for sociolinguistic and contact linguistics research by revealing which of several available languages in a speech community is used for which topic, in what discourse genre, and by speakers of what educational profile, among other issues.

The following section takes this issue one step further by considering how the information on language use thus gained can benefit language planning as well language maintenance and language revitalisation efforts, two core issues of applied aspects of language documentation.

3.4 Contributions to language planning and language maintenance and revitalisation

Standardisation, 'graphisation' (Fishman (1974)), orthography development and the creation of a written environment are important areas of language planning and contribute essentially to language maintenance and revitalisation efforts. The more materials resulting from these efforts reflect actual language use, the more they are likely to be read rather than turning into 'white elephants'. What constitutes actual language use and how it is best reflected in writing is by no means a trivial issue. Questions of scripts and orthographies notwithstanding, speech communities and linguists alike often have purist and prescriptive attitudes that result in the creation of idealised materials with very limited functionality. As observed by Dorian (1981, 1989, 1994) and Florey (2004), speakers of endangered languages tend to develop more linguistic purism as their language undergoes structural changes. Speakers of minority languages in multilingual contexts very often want to eliminate traces of language contact especially lexical matter from the contact language. Linguists who want to create a record of the 'true' language often inadvertently strengthen these purist attitudes. Even leaving aside the problematic issue of relying on speakers' intuitions when identifying contact influence,[6] language materials in which foreign influence is

[6] Thus, native speakers tend to recognise loanwords until they are totally integrated but generally do not recognise pattern borrowing, so that any 'cleaning up' is highly selective in any case.

edited out, e.g. through the replacement of loanwords with neologisms, are often unintelligible to their intended audience.

In addition, fieldworkers and native speakers alike are faced with problems when it comes to the identification of loanwords as opposed to code-switching. As argued in 3.3 above, knowledge about the linguistic profiles of speakers can help in resolving this: speakers who use material of foreign origin although they do not speak the language in question have integrated it into their lexicon. Use of foreign matter by a large number of speakers without proficiency in the donor language means those items should be regarded as nativised vocabulary.

Language materials can be even more hand-tailored to meet the needs of specific groups, taking into account differences in the educational backgrounds of women and men, for example. Thus, for Jalonke, brochures on health aimed mainly at women would ideally look different from those for men because the latter are familiar with and prefer French 'scientific' terms, whereas women are better served by Jalonke or Fula terms.

Similar observations hold for language standardisation – information contained in corpora on usage patterns favoured by different age or sex groups may influence decisions on varieties and styles to adopt when 'reducing a language to writing' (Pike (1959))and selecting genres and texts to be represented in written form.

To conclude, corpus data can provide detailed information on actual language use that is otherwise difficult to obtain, and can counterbalance inaccurate impressions from qualitative observations or tainted by purist attitudes of speakers and researchers. In view of the limited resources available for endangered and minority languages and of the urgency of taking measures to assist their maintenance and revitalisation, corpus data that provide valuable materials for speech communities should not be underrated.

4. Issues in the creation and use of field-based corpora

The following sections sketch a number of areas in which a methodological discussion is needed in order to make the expertise of specialist fields in linguistics available to field-based corpora.

4.1 Corpus size

Corpora of well-documented languages are enormous compared to the overwhelming majority of field-based corpora of lesser described languages. To cite but two examples, the Brown corpus of American English (http://helmer.aksis.uib.no/icame/brown/bcm.html) contains over a million words while the British National Corpus (http://www.natcorp.ox.ac.uk/) has 100 million words. Problems of sample size like the following need to be researched in order to inform the compilation of smaller corpora:

- What sample size is likely to be representative and give a reliable overview of at least high-frequency phenomena?

- On what sample size can nonparametric statistical tests be undertaken with an adequate degree of confidence to prove or disprove the significance of quantitative observations?

4.2 Corpus structure and cross-linguistic comparability

Closely related to the question of corpus size is the issue of the internal structure of field-based corpora. Large corpora can aim at a balanced sampling of genres, speakers/authors, and text types, since they generally can draw upon a large number of written sources and a huge speaker population from which to collect samples of spoken language. Field-based corpora are much more subject to practical limitations, often resulting in 'convenience' corpora that can only represent a very limited number of speakers with a small array of genres and texts. Optimising the internal structure of field-based corpora would increase their representativeness, and therefore a systematic investigation of the following questions would be beneficial:

- How many speakers are needed for a realistic picture of constant and varying patterns of language use?

- What is the minimal amount of data needed per speaker, genre, text, etc, in order to establish a reliable profile for the category of interest?

The internal composition of a corpus also has consequences for its usefulness in the cross-linguistic comparison of corpus data, for instance comparing linguistic features associated with certain genres/registers across languages (see Biber (1994); Givon 1979)). Within corpus linguistics, great efforts have been made to create comparable corpora, that is, corpora of distinct languages that are as equivalent in the genres and text types included as possible, in order to facilitate comparative studies. To make field-based corpora maximally suited for such typological research, an investigation of the following would be extremely helpful:

- What are the non-linguistic 'dimensions of register variation' (Biber (1995), see also Himmelmann (1998)), such as formality, politeness, plannedness, etc., likely to be reflected in differences in linguistic structure across languages?

- What are the genres that these patterns tend to be associated with?

- To what extent do differences in topic influence the characteristics of genres?

It is obvious that the questions raised above represent a methodological challenge in that more data from field-based corpora are needed in order to answer them even partly while at the same time any answers would have an influence on the size, structure, and kinds of investigations of these corpora. In view of the important contributions that

these corpora can make, however, it is certainly worthwhile to aim at transforming this vicious circle into an upward spiral.

5. Conclusion

This paper has taken corpus data resulting from fieldwork on Jalonke as an example to illustrate the usefulness of field-based corpora for different linguistic disciplines. I have exemplified how these data can contribute to linguistic theory building in a quantitative study on argument realisation. I have further shown how corpus data can reveal different syntactic patterns of relevance to genre and register studies and how sociolinguistic information can be obtained from them. Finally, I have discussed how information on actual language use is relevant for applied aspects of language documentation such as the creation of written materials for the speech community. In view of the focus of language documentation on amassing primary language data, it can be expected that the role of field-based corpora will grow over the next decades. Therefore, research on the optimisation of these corpora is in order so that they can reach their potential to be a reliable and quantifiable basis for present and future research as well as serving the speech communities.

6. References

Abdulaziz, Mohamed H. 2003. The history of language policy in Africa with reference to language choice in education. In Adama Ouane (ed.). *Towards a multilingual culture of education*. 181-99. Hamburg: UNESCO Institute of Education.

Allen, Shanley, Sotaro Kita and Asli Ozyurek (ms). *New Tomatoman.*

Alsina, Alex 1996. The role of argument structure in grammar: evidence from Romance. *CSLI lecture notes* 62. Stanford, California: CSLI Publications.

Ameka, Felix K., Carlien de Witte and David P. Wilkins 1999. Picture series for positional verbs: Eliciting the verbal component in locative descriptions. In David Wilkins (ed.). *'Manual' for the 1999 Field Season*. 48-56. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Bakhtin, M.M. 1999 (1986). The problem of speech genres. In Adam Jaworski and Nikolas Coupland (eds.). *The discourse reader*. 121-32. London/New York: Routledge.

Batibo, Herman 2005. *Language decline and death in Africa: causes, consequences and challenges.* Clevedon: Multilingual Matters.

Bencini, Giulia M. and Adele E. Goldberg 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language* 43: 640-51.

Biber, Douglas 1994a. An analytical framework for register studies. In Douglas Biber and Edward Finegan (eds.). *Sociolinguistic perspectives on register*. 31-56. New York/Oxford: Oxford University Press.

Biber, Douglas 1994b. Using register-diversified corpora for general language studies. In Susan Armstrong (ed.). *Using large corpora.* 179-201. Cambridge, MA./London: The MIT Press.

Biber, Douglas 1995. *Dimensions of register variation.* Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad and Randi Reppen 1998. *Corpus linguistics. Investigating language structure and use: Cambridge approaches to linguistics.* Cambridge: Cambridge University Press.

Bickel, Balthasar 2003. Referential density in discourse and syntactic typology. *Language* 79: 708-736.

Bird, Charles S. 1971. Oral art in the Mande. In Carleton T. Hodge (ed.). *Papers on the Manding*. 15-25. The Hague: Mouton.

Bohnemeyer, Jürgen (1998). *Time relations in discourse. Evidence from a comparative approach to Yukatek Maya*. Katholieke Universiteit: Ph.D. thesis.

Bohnemeyer, Jürgen and Martijn Caelen 1999. The ECOM clips: a stimulus for the linguistic coding of event complexity. In David Wilkins (ed.). *'Manual' for the 1999 Field Season*. 74-86. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Bohnemeyer, Jürgen, Sonja Eisenbeiss and Bhuvana Narasimhan 2001. Event Triads. In Stephen C. Levinson and Nick Enfield (eds.). *'Manual' for the field season 2001*. 101-15. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Bohnemeyer, Jürgen, Melissa Bowerman and Penelope Brown 2001. Cut and Break Clips. In Stephen C. Levinson and Nick Enfield (eds.). *'Manual' for the field season 2001*. 90-97. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Bowerman, Melissa and Eric Pederson 1993. Topological relations pictures. Eve Danziger and Deborah Hill (eds.). *Manual for the Space Stimuli Kit 1.2*. 40-50.

Nijmegen: Max Planck Institut für Psycholinguistik, Cognitive Anthropology Research Group.

Bresnan, Joan 2001. *Lexical-functional syntax*. Malden, Mass.: Blackwell.

Butt, Miriam and Wilhelm Geuder 1998. The projection of arguments: lexical and compositional factors. *CSLI lecture notes* 83. Stanford, California: CSLI Publications.

Camara, Sory 1976. *Gens de la parole. Essai sur la condition et le rôle des griots dans la societé Malinké.* Paris/La Haye: Mouton.

Chafe, Wallace L. (ed.) 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Advances in Discourse Processes, 3.* Norwood, NJ: ABLEX Publishing Corporation.

Comrie, Bernard 1988. Passive and voice. In Masayoshi Shibatani (ed.). *Passive and voice*. 9-23. Amsterdam/Philadelphia: John Benjamins.

Comrie, Bernard and Maria Polinsky (eds.) 1993. *Causatives and transitivity.* Amsterdam/Philadelphia: John Benjamins.

Conrad, David C. and Barbara E. Frank (eds.) 1995. *Status and identity in West Africa: Nyamakalaw of Mande.* Bloomington: Indiana University Press.

Creissels, Denis 1983. *Eléments de grammaire de la langue mandinka.* Grenoble: Publications de l'université des langues et lettres.

Creissels, Denis 1991. *Description des langues négro-africaines et théorie syntaxique.* Grenoble: ELLUG.

Dorian, Nancy C. 1981. *Language death: the life cycle of a Scottish Gaelic dialect.* Philadelphia: University of Pennsylvania Press.

Dorian, Nancy C. 1989. *Investigating obsolescence: studies in language contraction and death.* Cambridge; New York: Cambridge University Press.

Dorian, Nancy C. 1994. Stylistic variation in a language restricted to private-sphere use. In Douglas Biber and Edward Finegan (eds.). *Sociolinguistic perspectives on register*. 217-32. New York/Oxford: Oxford University Press.

Dumestre, Gérard 1994. *Le bambara du Mali: essai de description linguistique: Les documents de linguistique africaine.*

Ferguson, Charles M. 1994. Dialect, register, and genre: working assumptions about conventionalisation. In Douglas Biber and Edward Finegan (eds.). *Sociolinguistic perspectives on register*. 15-30. New York/Oxford: Oxford University Press.

Fillmore, Charles J. 1986. Pragmatically controlled zero anaphora. *BLS* 12: 95-107.

Fillmore, Charles J. and Paul Kay 1993. *Construction grammar coursebook. Course reader for Linguistics X20.* Berkeley: University of California, Berkeley.

Finegan, Edward and Douglas Biber 1994. Register and social dialect variation: an integrated approach. In Douglas Biber and Edward Finegan (eds.). *Sociolinguistic perspectives on register*. 315-47. New York/Oxford: Oxford University Press.

Fishman, Joshua A. (ed.) 1974. *Advances in language planning.* Den Haag/Paris: Mouton.

Florey, Margaret 2004. Countering purism: confronting the emergence of new varieties in a training program for community language workers. *Language Documentation and Description* 2.

Fried, Mirjam and Jan-Ola Östman (eds.) 2004. *Construction grammar in a cross-language perspective.* Amsterdam/Philadelphia: John Benjamins.

Givon, Talmy 1979. *Discourse and syntax.* New York: Academic Press.

Givón, Talmy 1990. Verb serialisation in Tok Pisin and Kalam: A comparative study of temporal packaging. In John W. M. Verhaar (ed.). *Melanesian Pidgin and Tok Pisin: Proceedings of the First International Conference of Pidgins and Creoles in Melanesia.* 19-55. Amsterdam and Philadelphia: John Benjamins.

Goldberg, Adele E. 1995. *Constructions: a construction grammar approach to argument structure: Cognitive theory of language and culture.* Chicago: University of Chicago Press.

Goldberg, Adele E. 1997. The relationship between verbs and constructions. In Marjolijn Verspoor, Kee Dong Lee and Eve Sweetser (eds.). *Lexical and syntactical constructions and the construction of meaning*. 383-98. Amsterdam/Philadelphia: John Benjamins.

Goldberg, Adele E. 1998. Patterns of experience in patterns of language. The new psychology of language. In Michael Tomasello (ed.). *Cognitive and functional approaches to language structure*. 203-20. Mahwah, New Jersey/London: Lawrence Erlbaum Associates.

Goldberg, Adele E. 1999. The emergence of the semantics of argument structure constructions. In Brian Mac Whinney (ed.). *The emergence of language*. 197-212. Mahwah, New Jersey/London: Lawrence Erlbaum Associates.

Goldberg, Adele E. 2002. Surface generalisations: an alternative to alternations. *Cognitive Linguistics* 14: 327-56.

Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Science* 7: 219-24.

Grimshaw, Jane 1990. *Argument structure.* Cambridge, Mass.: MIT Press.

Guerssel, Mohammed, Kenneth Hale, Mary Laughren, Beth Levin, and Josie White Eagle (1985). A cross-linguistic study of transitivity alternations. *CLS* 21: 48-63.

Hale, Kenneth L. and Samuel Jay Keyser 2002. *Prolegomenon to a theory of argument structure.* Linguistic inquiry monographs 39. Cambridge, Mass.: MIT Press.

Haspelmath, Marin 1990. The grammaticisation of passive morphology. *Studies in Language* 14: 25-72.

Haspelmath, Martin 1987. Transitivity alternations of the anticausative type. *AKUP* 5.

Haspelmath, Martin 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie and Maria Polinsky (eds.). *Causatives and transitivity.* 87-111. Amsterdam/Philadelphia: John Benjamins.

Hellwig, Birgit and Friederike Lüpke 2001. Caused positions. In Stephen C. Levinson and Nick Enfield (eds.). *'Manual' for the field season 2001*. 126-28. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Himmelmann, Nikolaus 1998. Documentary and descriptive linguistics. *Linguistics* 36: 161-95.

Huang, C.-T. James 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry* 15: 531-74.

Huang, C.T. James 1989. Pro-drop in Chinese: a generalised control theory. In O. Jaeggli and K. Safir (eds.). *The null subject parameter*. 185-214. Dordrecht: Kluwer.

Hymes, Dell 1972. Models of the interaction of language and social life. In John J. Gumperz and Dell Hymes (eds.). *Directions in Sociolinguistics. The ethnography of communication*. 35-71. New York: Holt, Rinehart and Winston.

Hymes, Dell 1974. Ways of speaking. In Richard Bauman and Joel Sherzer (eds.). *Explorations in the ethnography of speaking,* 433-51. Cambridge: Cambridge University Press.

Kita, Sotaro 1995. Enter/Exit Animation for Linguistic Elicitation. In David Wilkins (ed.). *'Manual' for Field Elicitation for the 1995 Field Season*. 13. Nijmegen: Max Planck Institut für Psycholinguistik, Cognitive Anthropology Research Group.

Levin, Beth 1993. *English verb classes and alternations. A preliminary investigation.* Chicago/London: The University of Chicago Press.

Levin, Beth 1995. Approaches to lexical semantic representation. In A. Zampolli, D. Walker and N. Calzolari (eds.). *Automating the lexicon*. 53-91. Oxford: Oxford University Press.

Levin, Beth and Steven Pinker 1992. *Lexical & conceptual semantics: Cognition special issues.* Cambridge, Mass.: Blackwell.

Levin, Beth and Malka Rappaport Hovav 2003. *The Dative Alternation revisited.* Paper presented at the Workshop on verb classes and alternations, University of Stuttgart, January 2003.

Li, Charles N. 1997. On zero anaphora. In Joan L. Bybee, John Haiman and Sandra A. Thompson (eds.). *Essays on language function and language type. Dedicated to T. Givón*. 275-300. Amsterdam/Philadelphia: John Benjamins.

Li, Charles N. and Sandra A. Thompson 1979. Third-person pronouns and zero-anaphora in Chinese discourse. In Talmy Givón (ed.). *Discourse and syntax*. 311-35. New York: Academic Press.

Lüpke, Friederike 2005. *A grammar of Jalonke argument structure*. Radboud Unviersiteit Nijmegen [Max Planck Series in Psycholinguistics 30]: PhD thesis.

Lüpke, Friederike (forthcoming 2006). It's a split, but is it unaccusativity? Two classes of intransitive verbs in Jalonke. *Studies in Language.*

Makoni, Sinfree, Nkonko Kamwangamalu and Centre for Advanced Studies of African Society 2000. *Language and institutions in Africa.* CASAS book series 5. Cape Town, South Africa: Centre for Advanced Studies of African Society.

Mel'cuk, Igor 1993. Voice: towards a rigorous definition. In Maria Polinsky and Bernard Comrie (eds.). *Causatives and transitivity.* 1-46. Amsterdam/Philadelphia: John Benjamins.

Mohammed Guerssel, Kenneth Hale, Mary Laughren, Beth Levin and Josie White Eagle 1985. A cross-linguistic study of transitivity alternations. *CLS* 21: 48-63.

Myers-Scotton, Carol 1993. *Social motivations for codeswitching: evidence from Africa.* Oxford: Clarendon.

Nathan, David and Peter Austin 2004. Reconceiving metadata: language documentation through thick and thin. *Language Documentation and Description* 2: 179-87.

Ochs, Elinor 1979. Planned and unplanned discourse. In Talmy Givón (ed.). *Discourse and syntax.* 51-80. New York: Academic Press.

Östman, Jan-Ola and Mirjam Fried (eds.) 2005. *Construction grammars: cognitive grounding and theoretical extensions.* Amsterdam/Philadelphia: John Benjamins.

Pike, Kenneth Lee 1959. *Phonemics: a technique for reducing languages to writing*. University of Michigan publications. Linguistics, vol. 3. Ann Arbor: University of Michigan Press.

Pinker, Steven 1989. *Learnability and cognition: the acquisition of argument structure: Learning, development, and conceptual change.* Cambridge, Mass.: MIT Press.

Pu, Ming-Ming 1997. Zero anaphora and grammatical relations in Mandarin. In Talmy Givón (ed.). *Grammatical relations. A functionalist perspective*. 281-321. Amsterdam: John Benjamins.

Rappaport Hovav, Malka and Beth Levin 1998. Building verb meanings. In Miriam Butt and Wilhelm Geuder (eds.). *The projection of arguments*. 97-134. Stanford: CSLI Publications.

Rappaport Hovav, Malka and Beth Levin 2002. Change of state verbs: implications for theories of argument projection. *BLS* 28.

Slobin, Dan I. 1993. Frog Story Procedures. In Eve Danzinger and Deborah Hill *Space Stimuli Kit 1.2.* 51-52. Nijmegen: Max Planck Institut für Psycholinguistik, Cognitive Anthropology Research Group.

Thomason, Sarah Grey 2001. *Language contact*. Washington, D.C.: Georgetown University Press.

Thompson, Sandra A. and Paul J. Hopper 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. In Joan L. Bybee and Paul J. Hopper (eds.). *Frequency and the emergence of linguistic structure*. 27-60. Amsterdam/Philadelphia: John Benjamins.

Tröbs, Holger 1998. *Funktionale Sprachbeschreibung des Jeli (West-Mande). Vol. 3: Mande languages and linguistics.* Köln: Rüdiger Köppe Verlag.

van Hout, Angeliek 1996. *Event semantics of verb frame alternations. A case study of Dutch and its acquisition.* Tilburg: Tilburg Dissertation in Language Studies.

van Hout, Angeliek 2000. Event semantics in the lexicon-syntax interface: Verb frame alternations in Dutch and their acquisition. In James Pustejovsky and Carol Tenny. *Events as grammatical objects. The converging perspectives of lexical semantics and syntax.* 239-81: CSLI Publications.

Wilkins, David and Eve Danzinger 1999. Table Top Route Descriptions [re-publication of two previous manual entries]. In David Wilkins (ed.). *'Manual' for the 1999 Field Season*. 116-30. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.

Wilkins, David P. 1999. The 1999 demonstrative questionnaire: "This" and "that" in comparative perspective. In David Wilkins (ed.). *'Manual' for the 1999 Field Season*. 1-24. Nijmegen: Max Planck Institut für Psycholinguistik, Language and Cognition Group.