Information Structure in African Languages: Corpora and Tools

Christian Chiarcos, Ines Fiedler, Mira Grubic, Andreas Haida, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes, Malte Zimmermann

Collaborative Research Centre ‘Information Structure’
Universität Potsdam, Germany & Humboldt Universität zu Berlin, Germany

March 31, 2009

Table of contents

Introduction to the work of the CRC IS

The Collaborative Research Centre ‘Information structure’ .

- 42 researchers
- 4 disciplines (Linguistics, Psychology, German Studies, African Studies)
- 15 projects
- 2 universities (Humboldt-University Berlin, University of Potsdam)
- Funded by the German Research Foundation
- Common goal: better understanding of information structure across languages

Introduction to the work of the CRC IS

The Collaborative Research Centre ‘Information structure’ .

- 42 researchers
- 4 disciplines (Linguistics, Psychology, German Studies, African Studies)
- 15 projects
- 2 universities (Humboldt-University Berlin, University of Potsdam)
- Funded by the German Research Foundation
- Common goal: better understanding of information structure across languages

What is Information Structure?

Information Structure

Information Structure is the structuring of linguistic information in order to optimize information transfer relative to the temporary communicative needs of interlocutors.

What is Information Structure?

The same information needs to be 'packaged' in different ways depending on the knowledge and goals of the speakers.

(1) a. I have a cat, and I had to bring my cat to the vet.
b. #I had to bring my cat to the vet, and I have a cat.

What is Information Structure?

The same information needs to be ‘packaged’ in different ways depending on the knowledge and goals of the speakers.

(2) a. I have a cat, and I had to bring my cat to the vet.

b. #I had to bring my cat to the vet, and I have a cat.

What is Information Structure?

Important concepts: Focus

Focus indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions.

(3) a. Clyde had to marry BERtha$_F$ in order to be eligible for the inheritance.

b. Clyde had to MARry$_F$ Bertha in order to be eligible for the inheritance.

What is Information Structure?

Important concepts: Focus

Focus indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions.

(4) a. Clyde had to marry BERtha$_F$ in order to be eligible for the inheritance.

b. Clyde had to MARry$_F$ Bertha in order to be eligible for the inheritance.

What is Information Structure?

(5) a. Who stole the cookie?

b. PEter$_F$ stole the cookie.

c. #Peter stole the COOkie$_F$.

What is Information Structure?

Important concepts: Givenness

Givenness is the indication that a concept is immediately present in the shared knowledge of the speakers, e.g. previously mentioned:

(6) a. Who stole the cookie?

b. PEter$_F$ [stole the cookie]$_{Given}$.

What is Information Structure?

Important concepts: Givenness

Givenness is the indication that a concept is immediately present in the shared knowledge of the speakers, e.g. previously mentioned:

(7) a. Who stole the cookie?

b. PEter$_F$ [stole the cookie]$_{Given}$.

What is Information Structure?

Important concepts: Givenness

(8) a. I know that John stole a cookie. What did he do then?

b. He [reTURNed [the cookie]$_{Given}$]$_F$

What is Information Structure?

Important concepts: Topic

The topic constituent identifies the entity under which the information expressed in the comment constituent should be ‘stored’.

(9) a. Aristotle Onassis$_{Topic}$ married Jacqueline Kennedy$_{Comment}$.

b. Jacqueline Kennedy$_{Topic}$ married Aristotle Onassis$_{Comment}$.

What is Information Structure?

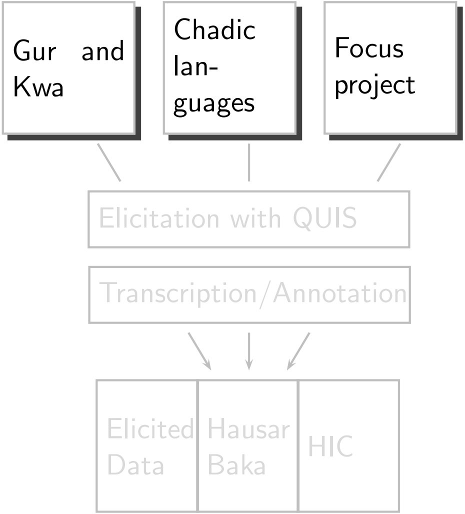
Important concepts: Topic

The topic constituent identifies the entity under which the information expressed in the comment constituent should be ‘stored’.

(10) a. Aristotle Onassis$_{Topic}$ married Jacqueline Kennedy$_{Comment}$.

b. Jacqueline Kennedy$_{Topic}$ married Aristotle Onassis$_{Comment}$.

Research at the CRC



Gur and Kwa

Chadic languages

Focus project

Elicitation with QUIS

Transcription/Annotation

Elicited Data

Hausar Baka

HIC

Information Structure in African Languages

- Focus marking by movement (Ex-situ focus)

(11)

Kiifii	nèe	Kande	ta-kèe	dafàa-waa.
fish	PRT	Kande	3sg-rel.cont	cook-NMLZ

(Hausa, Chadic)

‘Kande is cooking FISH.’

(12)

padgo	taabéè	**Kai**	(Tangale, Chadic)
bought	tobacco	Kai	

‘KAI bought tobacco.’

Information Structure in African Languages

- Focus marking without movement (In-situ focus)

(13)

púū	nʊ̄ndə́	ū	**bíí-gə̄**	*yə̀*	sábə̀-lə́.
woman	buy	CL.POSS	child-CL	FM	book-CL

(Byali, Gur)

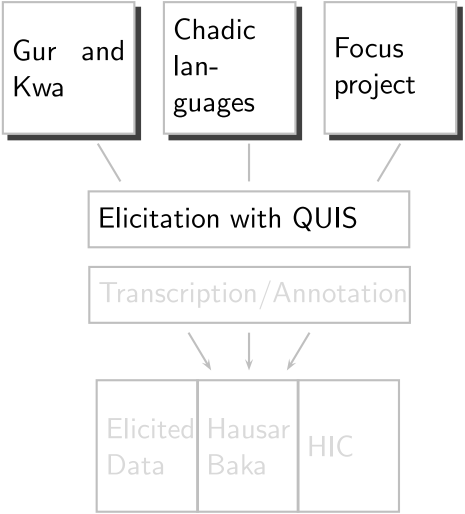
‘The woman bought a book for her CHILD.’

(14)

Yaa	sòokee	shì	dà	**wuƙaa.**	(Hausa, Chadic)
3sg.perf	stab	him	with	knife	

‘He stabbed him with a KNIFE.’

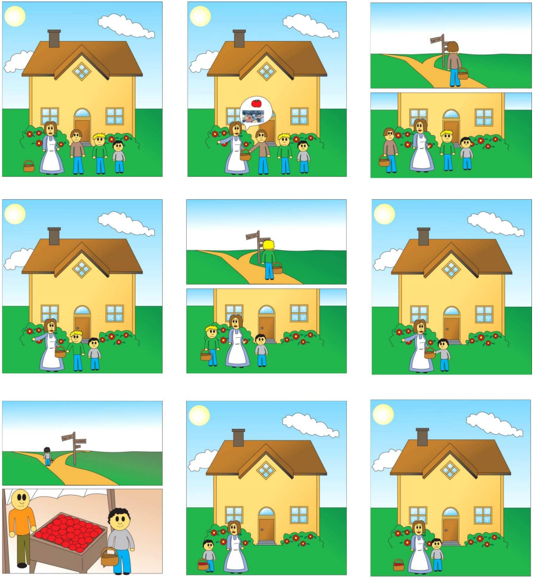
Research at the CRC



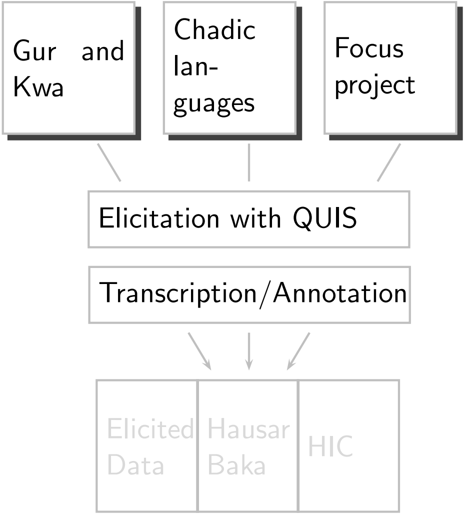
Questionnaire on IS

- (Skopeteas et al., 2006)
- Elicitation on the basis of pictures / short movies
- Descriptions, Narration, Questions/answers, Games
- highly controlled as well as less controlled settings

Questionnaire on IS



Research at the CRC



Gur and Kwa

Chadic languages

Focus project

Elicitation with QUIS

Transcription/Annotation

Elicited Data

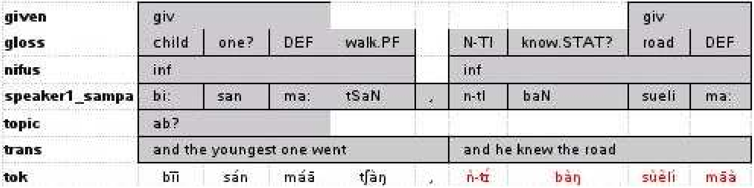
Hausar Baka

HIC

Transcription and Annotation

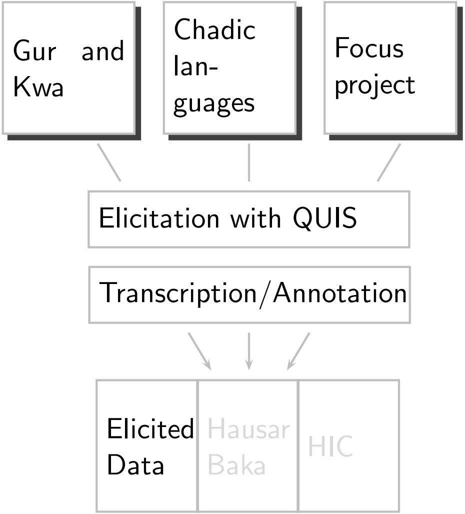
- annotation scheme LISA, (Dipper et al., 2007)
- applicable across typologically different languages
- guidelines for annotation of phonology, morphology, syntax, semantics and information structure
- (Semi-)automatic annotation also possible

Transcription and Annotation



given	giv								giv	
gloss	child	one?	DEF	walk.PF		N-TI	know.STAT?	road	DEF	
nifus	inf					inf				
speaker1_sampa	bi:	san	ma:	tSaN	,	n-tI	baN	sueli	ma:	
topic	ab?									
trans	and the youngest one went					and he knew the road				
tok	bīī	sán	máā	tʃàŋ	,	ǹ-tɪ́	bàŋ	sùèlí	māà	

Research at the CRC



Gur and Kwa

Chadic languages

Focus project

Elicitation with QUIS

Transcription/Annotation

Elicited Data

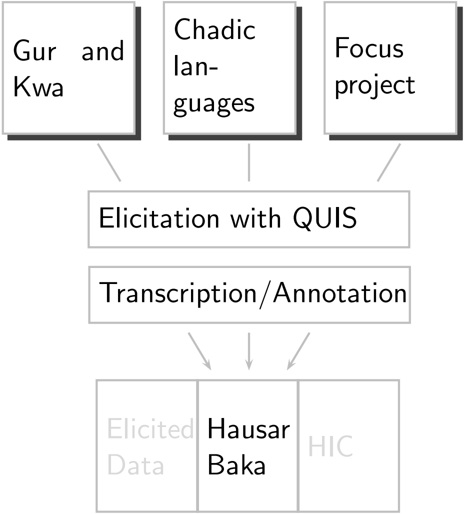
Hausar Baka

HIC

Elicited Data

- 19 Gur/Kwa languages: Baatonum, Buli, Byali, Dagbani, Ditammari, Gurene, Konkomba, Konni, Nateni, Waama, Yom (Gur languages) and Aja, Akan, Efutu, Ewe, Fon, Foodo, Lelemi, Anii (Kwa languages).
- 6 Chadic languages: Hausa, Tangale, Guruntum (West Chadic) and Bura, South Marghi, Tera (Central Chadic).
- elicited with QUIS and language-specific additional tasks.

Research at the CRC



Gur and Kwa

Chadic languages

Focus project

Elicitation with QUIS

Transcription/Annotation

Elicited Data

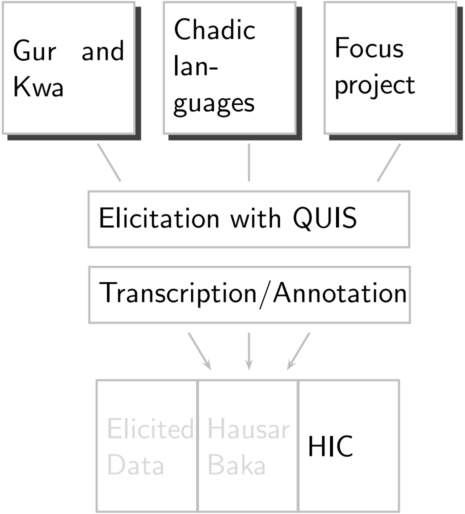
Hausar Baka

HIC

Hausar Baka Corpus

- by Randell, Bature and Schuh, 1998
- collection of videotaped dialogues
- about 1500 Hausa sentences
- annotated using LISA

Research at the CRC



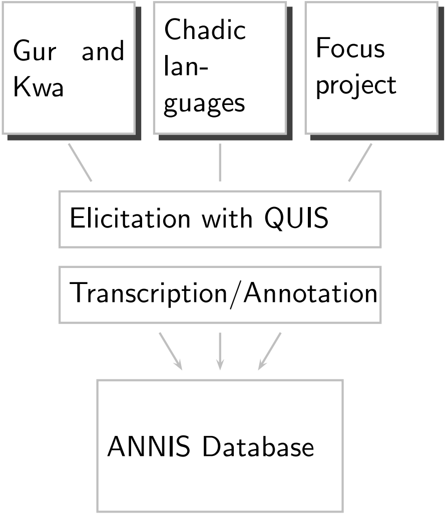
Hausa Internet Corpus

- current project
- in cooperation with another NLP project of the CRC
- large amounts of Hausa material available on the internet
- parallel sections: novel Ruwan Bagaja by Abubakar Imam, Bible and Qur'an sections, Declaration of Human Rights.
- These parallel sections open the possibility of semiautomatic annotation:
- POS annotation projection from English to Hausa
- Projected annotation used to train tagger/chunker
- Existing manual annotations used as a gold standard for evaluation

Hausa Internet Corpus

- current project
- in cooperation with another NLP project of the CRC
- large amounts of Hausa material available on the internet
- parallel sections: novel Ruwan Bagaja by Abubakar Imam, Bible and Qur'an sections, Declaration of Human Rights.
- These parallel sections open the possibility of semiautomatic annotation:
- POS annotation projection from English to Hausa
- Projected annotation used to train tagger/chunker
- Existing manual annotations used as a gold standard for evaluation

Research at the CRC



Gur and Kwa

Chadic languages

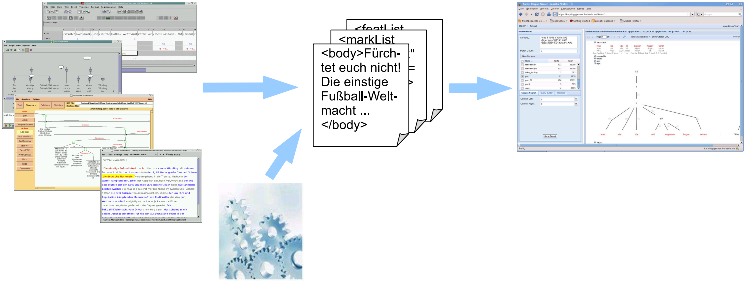
Focus project

Elicitation with QUIS

Transcription/Annotation

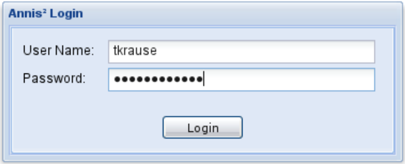
ANNIS Database

Framework Architecture


<featList
<markList
<body>Fürch-tet euch nicht! Die einstige Fußball-Welt-macht ... </body>

ANNIS

- web-based corpus interface
- query and visualization of annotations
  - (sequences of) tokens
  - trees (labeled edges, crossing edges)
  - pointing relations
  - nested, overlapping, conflicting, discontinuous
- user management
  - authorized access
  - acc. to legal status of corpus



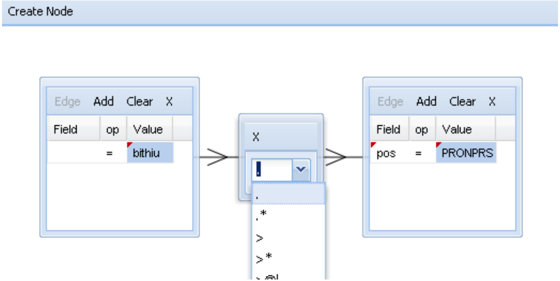
Annis^2 Login

User Name: tkrause

Password: ●●●●●●●●●●●●●

Querying in ANNIS

- ANNIS Query Language
- graphical Query Builder (drag & drop)



Field	op	Value
	=	bithiu

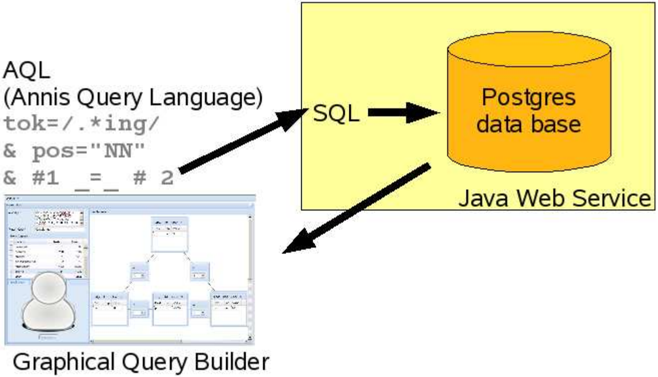
Field	op	Value
pos	=	PRONPRS

basic concepts:
nodes, relations between nodes

ANNIS Query Language

- nodes (sequentially numbered variables)
  - generalized category
    `tok` (= any token), `node` (= any annotation)
  - regular expressions / exact expressions
    `pos=/ADJ[AD]/, pos=/P.*/, cat="NP"`
- relations between nodes
  - co-extension, overlapping, contained/adjacent span
    `lemma=/.*ing/ & pos="NN" & #1 _=_ #2`
  - dominance (direct/indirect, left-/rightmost child, common parent, etc., including edge labels)
    `cat="NP" & cat="PP" & #1 > #2`

Query Processing



AQL
(Annis Query Language)

```
tok=/.*ing/
& pos="NN"
& #1 _=_ # 2
```

SQL

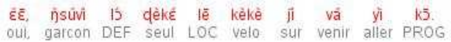
Postgres data base

Java Web Service

Graphical Query Builder

Corpus Presentation

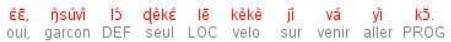
- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)



ɛ́ɛ̄,	ŋ̀súvì	lɔ́	ɖèkɛ́	lē	kèkè	jī	vá	yì	kɔ̄.
oui,	garcon	DEF	seul	LOC	velo	sur	venir	aller	PROG

Corpus Presentation

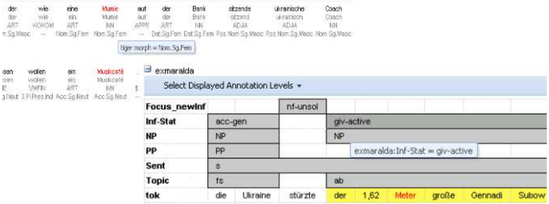
- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)



ɛ̌ɛ̄,	ŋ̀súvì	lɔ́	ɖèkɛ́	lē	kèkè	jī	vǎ	yì	kɔ̄.
oui,	garcon	DEF	seul	LOC	velo	sur	venir	aller	PROG

Corpus Presentation

- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)
- visualization of annotations
  - tokens, spans



der	wie	eine	Mumie	auf	der	Bank	sitzende	ukrainische	Coach
der	wie	ein	Mumie	auf	der	Bank	sitzend	ukrainisch	Coach
ART	KOKOM	ART	NN	APPR	ART	NN	ADJA	ADJA	NN
n.Sg.Masc	--	Nom.Sg.Fem	Nom.Sg.Fem	--	Dat.Sg.Fem	Dat.Sg.Fem	Pos.Nom.Sg.Masc	Pos.Nom.Sg.Masc	Nom.Sg.Masc

tiger:morph = Nom.Sg.Fem

sen	wollen	ein	Musikcafé
sen	wollen	ein	Musikcafé
E	VMFIN	ART	NN
g.Neut	3.Pl.Pres.Ind	Acc.Sg.Neut	Acc.Sg.Neut

exmaralda

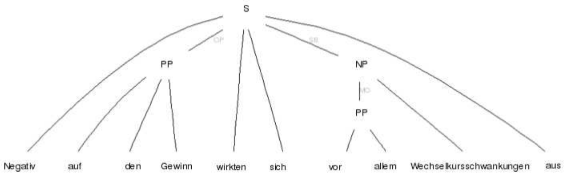
Select Displayed Annotation Levels

Focus_newInf			nf-unsol						
Inf-Stat	acc-gen			giv-active					
NP	NP			NP					
PP	PP								
Sent	s								
Topic	fs			ab					
tok	die	Ukraine	stürzte	der	1,62	Meter	große	Gennadi	Subow

exmaralda:Inf-Stat = giv-active

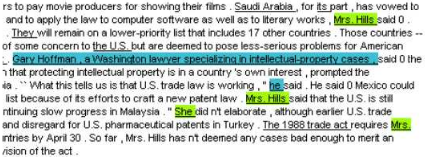
Corpus Presentation

- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)
- visualization of annotations
  - tokens, spans
  - trees



Corpus Presentation

- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)
- visualization of annotations
  - tokens, spans
  - trees
  - pointing relations



rs to pay movie producers for showing their films . Saudi Arabia , for its part , has vowed to
and to apply the law to computer software as well as to literary works , Mrs. Hills said 0 .
. They will remain on a lower-priority list that includes 17 other countries . Those countries --
of some concern to the U.S. but are deemed to pose less-serious problems for American
. Gary Hoffman , a Washington lawyer specializing in intellectual-property cases , said 0 the
1 that protecting intellectual property is in a country 's own interest , prompted the
ia . `` What this tells us is that U.S. trade law is working , " he said . He said 0 Mexico could
list because of its efforts to craft a new patent law . Mrs. Hills said that the U.S. is still
ntinuing slow progress in Malaysia . " She did n't elaborate , although earlier U.S. trade
and disregard for U.S. pharmaceutical patents in Turkey . The 1988 trade act requires Mrs.
ıntries by April 30 . So far , Mrs. Hills has n't deemed any cases bad enough to merit an
/ision of the act .

Corpus Presentation

- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)
- visualization of annotations
  - tokens, spans
  - trees
  - pointing relations
- rendering of audio files (embedded media player)
- save and export facilities
  - 'deep links' for citation
  - export to tabular format ARFF (WEKA machine learning environment)

Corpus Presentation

- match count for quantitative studies
- full Unicode support (diacritics, e.g. for tone)
- visualization of annotations
  - tokens, spans
  - trees
  - pointing relations
- rendering of audio files (embedded media player)
- save and export facilities
  - 'deep links' for citation
  - export to tabular format ARFF
    (WEKA machine learning environment)

Summary

- Resources
  - deeply annotated
  - specialized on IS
  - tools allowing for query and evaluation
- extend corpus studies
  - near-natural language
  - larger amounts of data
- better understanding of IS