
Workshop on multilingual data, 08 July 2003
MULTILINGUAL DATABASE: Obstacles and Opportunities
 Thomas Schmidt, Project Zb

Data at the SFB "Mehrsprachigkeit"

K1: Japanese and German expert discourse in mono- and multilingual settings (JadEx)

Discourse type:	Expert discourse
Languages:	Japanese, German
Background:	Discourse Analysis (Functional Pragmatics)
Transcription system:	HIAT
Software:	syncWriter (Macintosh)

K2: Interpreting in hospitals

Discourse type:	Interpreter mediated doctor-patient communication
Languages:	Portuguese, Turkish, German
Background:	Discourse Analysis (Functional Pragmatics)
Transcription system:	HIAT
Software:	syncWriter (Macintosh)

K4: Covert Translation

Text types:	Parallel and translated texts from popular science, business, computer instructions
Languages:	English, German, French, Spanish
Background:	Model based on systemic functional language theory, speech act theory, discourse analyses
Software:	Text editors, Corpus processing software (Concordancer, Parallel text aligner, Systemic coder) (Windows)

K5: Semicommunication and receptive multilingualism in contemporary Scandinavia

Discourse type:	Radio broadcasts, Group discussions, School lessons
Languages:	Danish, Swedish, Norwegian
Background:	Discourse Analysis
Transcription system:	HIAT
Software:	HIAT-DOS (Windows)

E1: Bilingualism in early childhood: Comparing Italian/German and French/German

Discourse type:	Child/Interviewer discourse
Languages:	Italian, French, German
Background:	Generative Syntax

E2: Simultaneous and successive acquisition of bilingualism

Discourse type:	Child/Interviewer discourse
Languages:	French, Basque, Portuguese, German
Background:	Generative Syntax
Transcription system:	Project's own conventions
Software:	LAPSUS / dBASE III / dBASE IV (Windows)

E3: Prosodic constraints on phonological and morphological development in bilingual first language acquisition

Discourse type: Child/Interviewer discourse
 Languages: Spanish, German
 Background: Generative Phonology / Optimality theory
 Transcription system: Project's own conventions
 Software: 4th dimension / SoundScope / Praat (Macintosh)

E4: Specific Language Impairment and early L2 Acquisition: Differences in Grammatical Development

Discourse type: Child/Interviewer discourse
 Languages: German, Turkish
 Background: Generative Syntax
 Transcription system: Project's own conventions
 Software: Project's own video transcription software / annotation software / analysis software (Macintosh)

E5: Linguistic connectivity in bilingual Turkish-German children (SKOBI)

Discourse type: Child/Interviewer discourse, Family communication, Story retelling
 Languages: Turkish, German
 Background: Discourse Analysis (Functional Pragmatics)
 Transcription system: HIAT
 Software: syncWriter (Macintosh)

H1: Multilingualism as cause and effect of language change: Historical syntax of Roman languages

Text type: Chronicles, Historical Documents
 Languages: Old French / Occitan / Portuguese
 Background: Generative syntax
 Software: ACCESS (Windows)

H3: Scandinavian syntax in multilingual context

Text type: Narrative texts, Bibel translations, Juridical texts, Historical records
 Languages: Ancient Nordic, Latin, Old Swedish, Old Danish, Middle Low German, New High German
 Background: Syntactic analysis
 Software: ACCESS (Windows)

H4: Forms of written discourse in Byzantine and Modern Greek Diglossia

Text type: Narrative texts (intralingual translation)
 Languages: Different H- and L-variants of Greek
 Background: Discourse Analysis / Systemic functional linguistics
 Software: Text editors, Concordancers, ACCESS, Systemic Coder (Windows & Macintosh)

H5: Multilingualism, Linguistic Variation, and Linguistic Universals

Text type: Irish Emigrants' letters
 Languages: Older forms of Irish English (18th and 19th century)
 Software: ACCESS, Text editors (Windows)

E5: Linguistic connectivity in bilingual Turkish-German children (SKOBI)

Sel [v]	We nn du	Spanish willst, musst	Du	Spanish lernen oder Fransösüs	oder wenn du	((1s)) Dings...				
Sel [tl]	COMP	PRON2SGSpanish	want:PRES:2SGmust:PRES:2SGPron2SGSpanish	learn:INF	CONJ	French	CONJ	COMP	PRON2SG	thing
Sel [ts]	When you want Spanish, you have to learn Spanish or French				or when you ((1s)) how do you call it...					
Yil [v]	Nasıl?								Benimle lütfen	
Yil [ts]	How?								Would you mind	
Sel [k]	<i>für: Französisch</i>									

E2: Simultaneous and successive acquisition of bilingualism

	CHILD	INTERVIEWER	COMMENT
1	hier du muß hier das hinstell'n U D M U D V		Kind zeigt Interviewerin, wo sie die Spielfiguren hinstellen muss.
2	gib mir meine roten erstmal dann bin ich dein freund V D D A U U I D D N		
3	dann stell' ich die hin aber U V D D U U	da sind deine roten	Interviewer und Y lachen, Interviewer gibt Kind die roten Spielfiguren
4	mußt du so drehn o.k. M D U V U		Kind zeigt Interviewer, wie man das Krokodil drehen kann
7	wenn du hier machst dann muß du auf den krokodil dreh'n C D U V U M D P D N V		Kind zeigt auf die Würfelseite mit dem Krokodilkopf

Diversity and points of comparison

Diversity in the data on several dimensions

- Spoken vs. written language
 - Written texts vs. Transcriptions of spoken language
- Different discourse types / Different text types
 - Multiparty discourse (Group discussions) vs. Monologues (Story retelling)
 - Journal articles vs. Personal letters
- Different languages / Different writing systems
 - Different extensions of the Latin alphabet (Portuguese / Turkish / Danish letters with diacritics, German Umlauts etc.)
 - Non-Latin alphabets (Greek, Runic)
 - Non-alphabetic writing systems (Japanese)
- Different theoretical background / Different research goals / Different methodologies
 - Functional pragmatics vs. Generative syntax
 - Syntax vs. Phonology
 - Synchronic vs. diachronic research
 - Qualitative vs. quantitative methods
- Different transcription systems
 - Different conventions (HIAT vs. E2 conventions)
 - Different selection of things to be transcribed / different level of detail in transcription (overlap, non-verbal behaviour, "Performance phenomena", orthographic vs. phonologic transcription)
 - Different notational systems (Partitur vs. column notation)
 - Different media (Audio vs. Video data)
- Different Software
 - Different operating systems (Windows vs. Mac OS)
 - Different tool types (Partitur editors vs. Databases vs. Text editors etc.)
 - Different tools (syncWriter vs. HIAT-DOS, dBase IV vs. 4th Dimension)
 - Different file formats (Binary vs. Text vs. XML)
 - Different character encodings (MacRoman vs. ANSI-1252 vs. JIS vs. UTF-8)

Interrelatedness of dimensions

- The choice of a transcription system depends on the theoretical background and the research goal (Ochs 1979: "Transcription as theory")
- The choice of a transcription software depends on the choice of the transcription system (Partitur editor for HIAT transcriptions)
- The choice of an operating system depends on the choice of a transcription software or vice versa (Macintosh for syncWriter / HIAT-DOS for Windows, 4th Dimension for Macintosh, dBASE for Windows)

Points of comparison

- Multilingual data (all projects)
- Transcription data (K1, K2, K5, E1, E2, E3, E4 and E5)
- Written data (K4, H1, H3, H4, H5)
- Historical texts (H-projects)
- Acquisition data (E-projects)
- German/Romance bilinguals (E1, E2 and E3)
- Turkish/German bilinguals (E4 and E5)
- Indo-European/Non-Indo-European bilinguals (E2, E4 and E5)
- HIAT data (K1, K2, K5, E5)
- English texts (K4 and H5)
- "Expert discourse" (K1, K2, K4, K5)

Obstacles and Opportunities

Sharing data: Opportunities

- Understand other people's analyses
- Verify or falsify other people's analyses
- Test own hypotheses on other people's data
- Use other people's tools
- Get statistically sound results (Representativeness)
- Keep data usable (exchange in time)

Sharing data: Obstacles

- Linguistic diversity (different languages, text/discourse types, written vs. spoken)
- Theoretical diversity (methodology, transcription systems, research goals)
- Technical diversity (tools, file formats, operating systems)
- Legal aspects
- Scientific competition
- Priorities: time and effort