

Intl' DH Congress 2012
2012-07-19
Hamburg



Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Sebastian Drude, Paul Trilsbeek, Daan Broeder

The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands



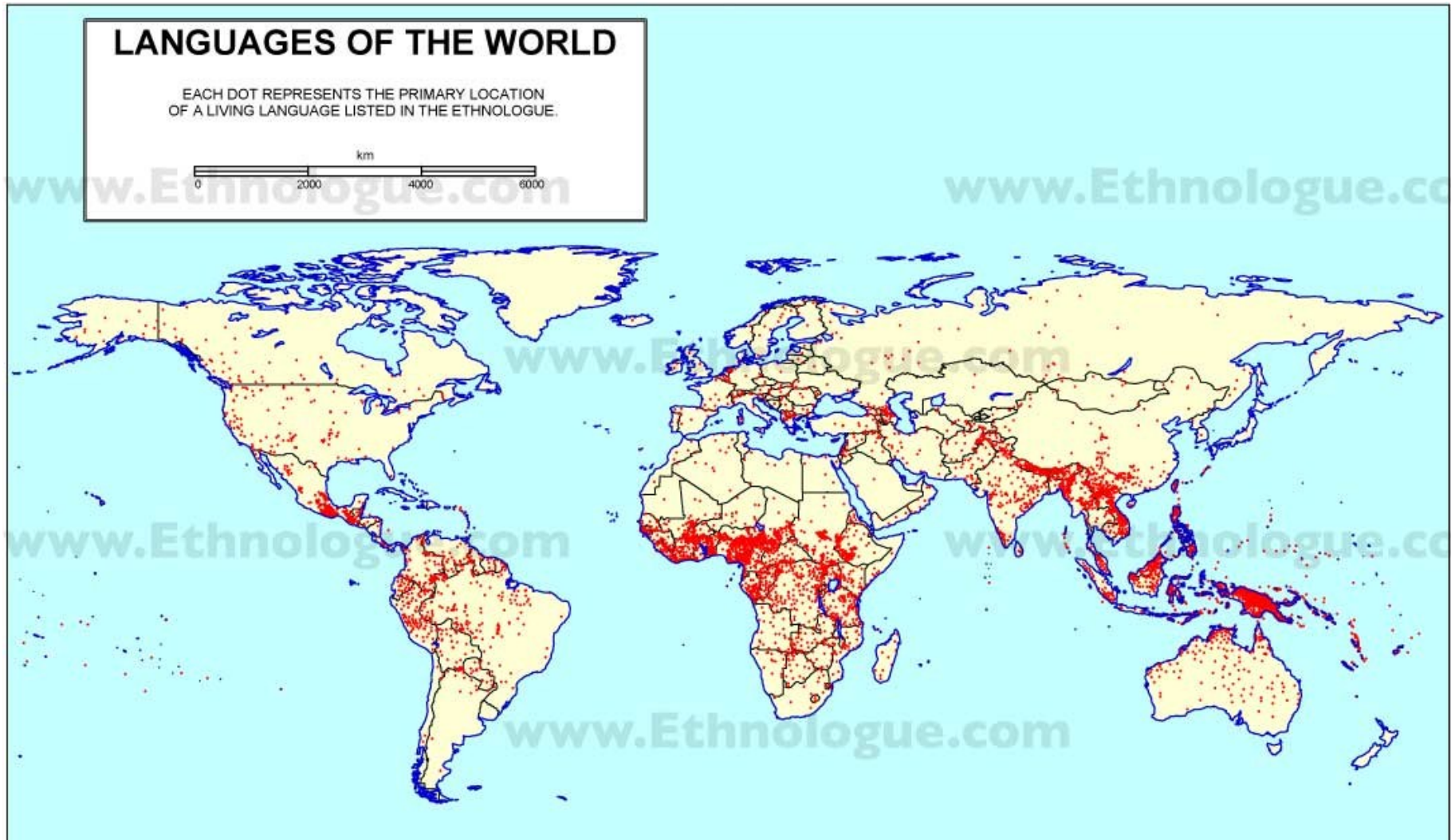
- 1) Language diversity and endangerment
- 2) Language Documentation
- 3) The DOBES Programme
- 4) The Language Archive
- 5) Relation to the Digital Humanities



- 1) **Language diversity and endangerment**
- 2) Language Documentation
- 3) The DOBES Programme
- 4) The Language Archive
- 5) Relation to the Digital Humanities








1) Linguistic diversity





1) Language endangerment



Degree of endangerment	Intergenerational Language Transmission
<i>safe</i>	language is spoken by all generations; intergenerational transmission is uninterrupted >> <i>not included in the Atlas</i>
 <i>vulnerable</i>	most children speak the language, but it may be restricted to certain domains (e.g., home)
 <i>definitely endangered</i>	children no longer learn the language as mother tongue in the home
 <i>severely endangered</i>	language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
 <i>critically endangered</i>	the youngest speakers are grandparents and older, and they speak the language partially and infrequently
 <i>extinct</i>	there are no speakers left >> <i>included in the Atlas if presumably extinct since the 1950s</i>

From: [UNESCO Atlas of the World's Languages in danger](#)



1) Language endangerment



- Aikana
- Ajuru
- Akawaio (Brazil)
- Akuntsu
- Akwáwa
- Amanayé
- Anambé
- Apalai
- Apiaká
- Apinajé
- Apurinã
- Arapáso
- Arara do Pará
- Arára
- Shawādáwa
- Araweté
- Arikapu
- Aruá
- Ashaninka (Brazil)
- Asurini do Xingu
- Aurê-Aurá
- Ava-Canoeiro
- Aweti
- Bakairi
- Banawá Yafi
- Baniwa do Içana
- Bara (Brazil)
- Barasana (Brazil)
- Baré (Brazil)
- Bororo



- 1) Language diversity and endangerment
- 2) **Language Documentation**
- 3) The DOBES Programme
- 4) The Language Archive
- 5) Relation to the Digital Humanities



2) Language documentation

- New subfield of linguistics (Himmelman 1998)
- Triggered by language endangerment, enabled by technical / digital revolution
- In addition to the “Boas’ian triad” (grammar, dictionary, text collection): **corpora of annotated multimedia-data**
- Focus on **natural language use**
- **Multi-purpose:**
 - language typology, linguistic analysis, ...
 - anthropology, ethno-musicology, ethno-biology...
 - oral history, archeology, lang. revitalization,...
- Some future uses may be unknown



2) Language documentation



THE HANS RAUSING
Endangered Languages Project



Documenting Endangered Languages (DEL)
data, infrastructure and computational methods





2) Language documentation



Essentials of Language Documentation

Edited by Jost Gippert, Nikolaus P. Himmelmann, Ulrike Mosel

mouton textbook

Mouton de Gruyter

LANGUAGE DOCUMENTATION & CONSERVATION

**THE MELD SCHOOL OF
Endangered Languages Project**
Because every last word means another lost world...

Language Documentation and Description
Volume 10

Edited by
Jan-Olof Svantesson,
Niclas Burenhult,
Arthur Holmer,
Anastasia Karlsson,
& Håkan Lundström

School of Oriental and African Studies

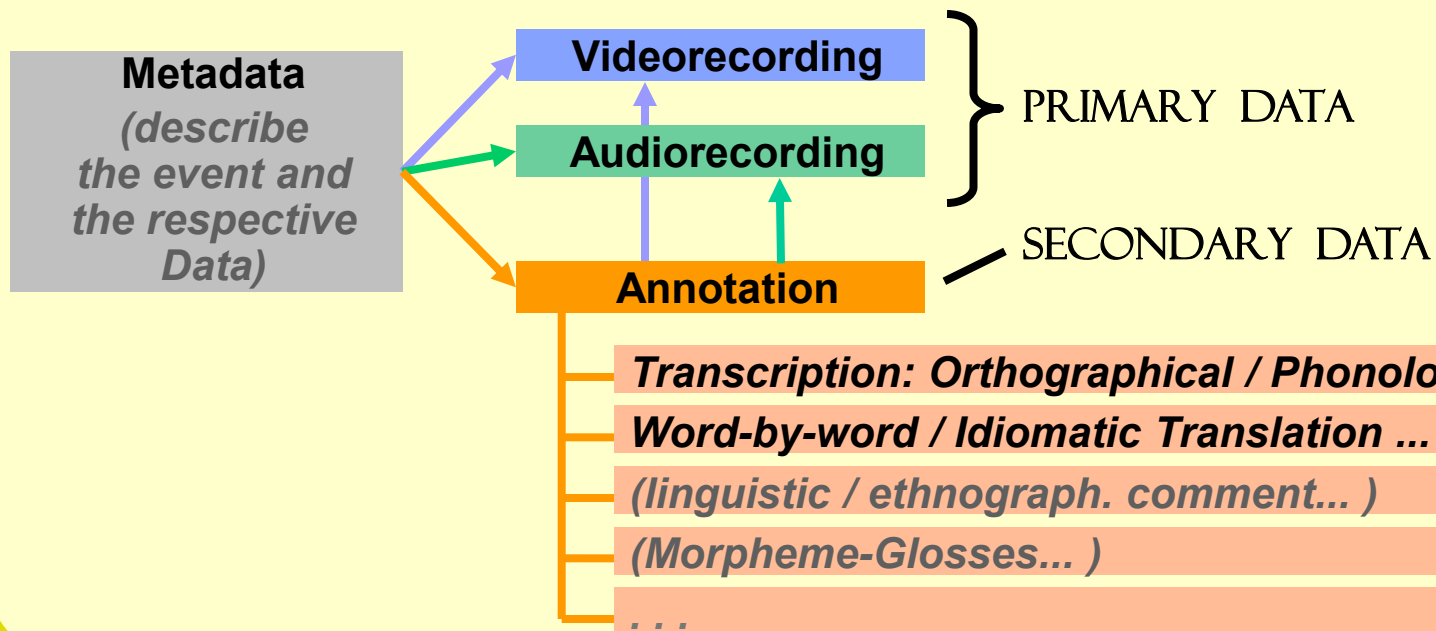
rnld
Resource Network for Linguistic Diversity

EMELD The E-MELD School of
Best Practices in Digital Language Documentation



2) Language documentation

SESSION





2) Language documentation



ELAN 3.9.1 File Edit Annotation Tier Type Search View Options Window Help

Elan - pear story.eaf

Grid Text Subtitles **Audio Recognizer** Video Recognizer Metadata Controls

Recognizer: Tag vowels (volume peaks of voiced timespans) File(s): pear.wav

Selections: Add Remove Add Tier...

Parameters: Pitch ceiling [Hz], eg. 300 male / 550 female (150.0 - 1500.0) 604.8
Intensity change [dB] to start/end a peak (0.5 - 5.0) 2.0
Minimum amplitude (0..1) for pitch analysis (0.02 - 0.3)

Segmentations: Create Tier(s)...

Progress: Ready Start Report...

00:00:16.840 Selection: 00:00:00.000 - 00:00:00.000 0

#Intensity[B] 78.005 55.3015 32.598
#pitch[Hz] 604.475 340.3425 76.21

38.9515 504.576

00 00:00:09.000 00:00:10.000 00:00:11.000 00:00:12.000 00:00:13.000 00:00:14.000 00:00:15.000 00:00:16.000 00:00:17.000 00:00:18.000 00:00:19.000 00:00:20.000

Event [1]
Clause Transcri [16] and so he climbs up a tree and he starts with the ladder and he starts picking pears off the tree and he puts the pears into an apron
Motion [16] motion non-motion non-motion
Gesture # [23] gestu gestu gesture 4 gestur gesture 6 gesture 7 gesture gesture 9 gesture 1
Gs Hand [23] R R R R R B B B



2) Language documentation

connections [0]			
A_po [1]	so it starts out with a rooster crows		
A_dt [1]	so it starts oooooout with ... a rooster crows		
A_mb [10]	so it start -s out with a rooster crow -s		
A_gl [10]	así lo empezar -3.sg fuera con un gallo cantar -3.sg		
A_tl [1]	así lo comienza (afuera) con un gallo canta		
A_tf [1]	entonces, se inicia con un canto del gallo		
A_tn [1]	então, começa com um canto de galo		
A_vb [1]	... gestos ...		
B_po			




2) Language documentation



- Text
- Grid
- Subtitle
- Waveform
- Timeline
- Combined

Video display min



Progress bar and playback controls: Play, Stop, Previous, Next, Full, Buffer.

Media information min

Resource: 010_autobiogr.eaf
 Media file: 010_autobiogr2.m4a

Elapsed time: 00:00:00:000

Selected chunk:
 Begin time: 00:00:00:000
 End time: 00:00:02:317
 Text: it was like that, son

Mini Data Frame min

it was like that, son here they were still lovers one day people saw them father got married mother got married mother had a child I grew in my mother's belly ready, the belly grew very big she brought me there to Japiaja, they brought me to the place my father's family was

Tier: SmngE@010
 Font size: 14

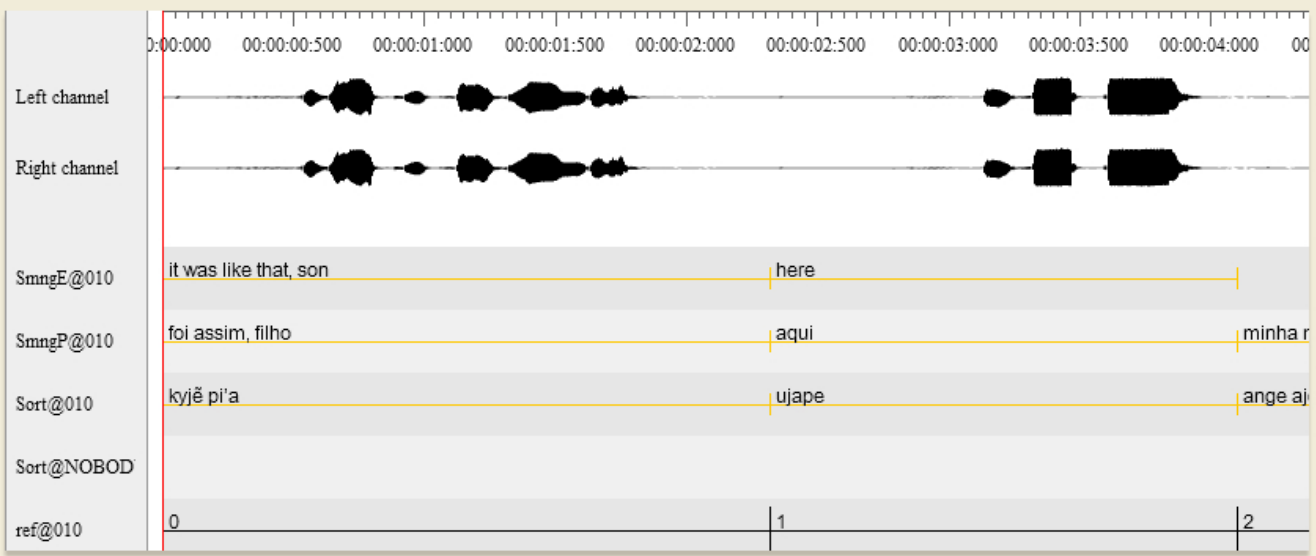
Play selection
 Clear selection

|< >|
 << >>
 < >
 + -

- Play screen by screen
- Play continually

Tier text font:
 Arial Unicode MS

Waveform and Timeline



Timeline segments:

SmngE@010	it was like that, son	here	
SmngP@010	foi assim, filho	aqui	minha r
Sort@010	kyjě pi'a	ujape	ange ai
Sort@NOBOD			
ref@010	0	1	2



- 1) Language diversity and endangerment
- 2) Language Documentation
- 3) **The DOBES Programme**
- 4) The Language Archive
- 5) Relation to the Digital Humanities



3) The DOBES programme



General:

- Initiative by the VolkswagenStiftung together with German linguists
- A programme rather than a project: a self-organized research structure
- Independent research teams, steering committee, advisory boards
- The heart is one central technical project and archive at the MPI Nijmegen, now “TLA”
- Each project 2-4 years on one or more languages



3) The DOBES programme



History:

- Interest in topic “language endangerment” since 1991/92, LSA symposium etc.
- Summer-school in Köln in 1993
- Conceptional preparation (until) 1999
- Pilot Phase 2000/2001:
5+3 individual projects & ‘TIDEL’ (TG @ MPI-PL)
- 2001-2012: calls for main phase,
each year around 7 new projects
- Total of ca. 70 individual projects (28Mio €)
on about 90 target languages



3) The DOBES programme





3) The DOBES programme



Achievements:

- General goals and methods have been established: that/how digital archives with primary data and annotation are built
- Relevant Software has been and is being developed (LAT: ELAN/ANNEX, LEXUS, IMDI/ARBIL, LAMUS/AMS)
- Legal and ethical questions have been addressed and partially answered
- National and regional archives are implemented
- Awareness of (endangerment of) language diversity
- LD established as field, other funding programs



- 1) Language diversity and endangerment
- 2) Language Documentation
- 3) The DOBES Programme
- 4) **The Language Archive**
- 5) Relation to the Digital Humanities



4) The Language Archive

- New unit at the Max-Planck-Institute for Psycholinguistics in Nijmegen, Netherlands





4) The Language Archive

- New unit at the Max-Planck-Institute for Psycholinguistics in Nijmegen, Netherlands
- Director: **W. KLEIN**, Head: **PETER WITTENBURG**
- MPG, MPI-PL, BBAW, KNAW contribute
- Goal: give long-term stability for archive, software development and other activities
- One of the backbone centers of CLARIN
- Participation in many international projects:
ISLE, MUMIS, ECHO, INTERA, DAM-LR, **DoBeS**, CGN, **CLARIN EU**, **CLARIN NL**, **CLARIN D**, **CLARA**, Inter, HARVE, **AVATech**, REPLIX, **RELISH**, **EUDAT**, **DASISH**, **INNET**, **iCORDI**



4) The Language Archive

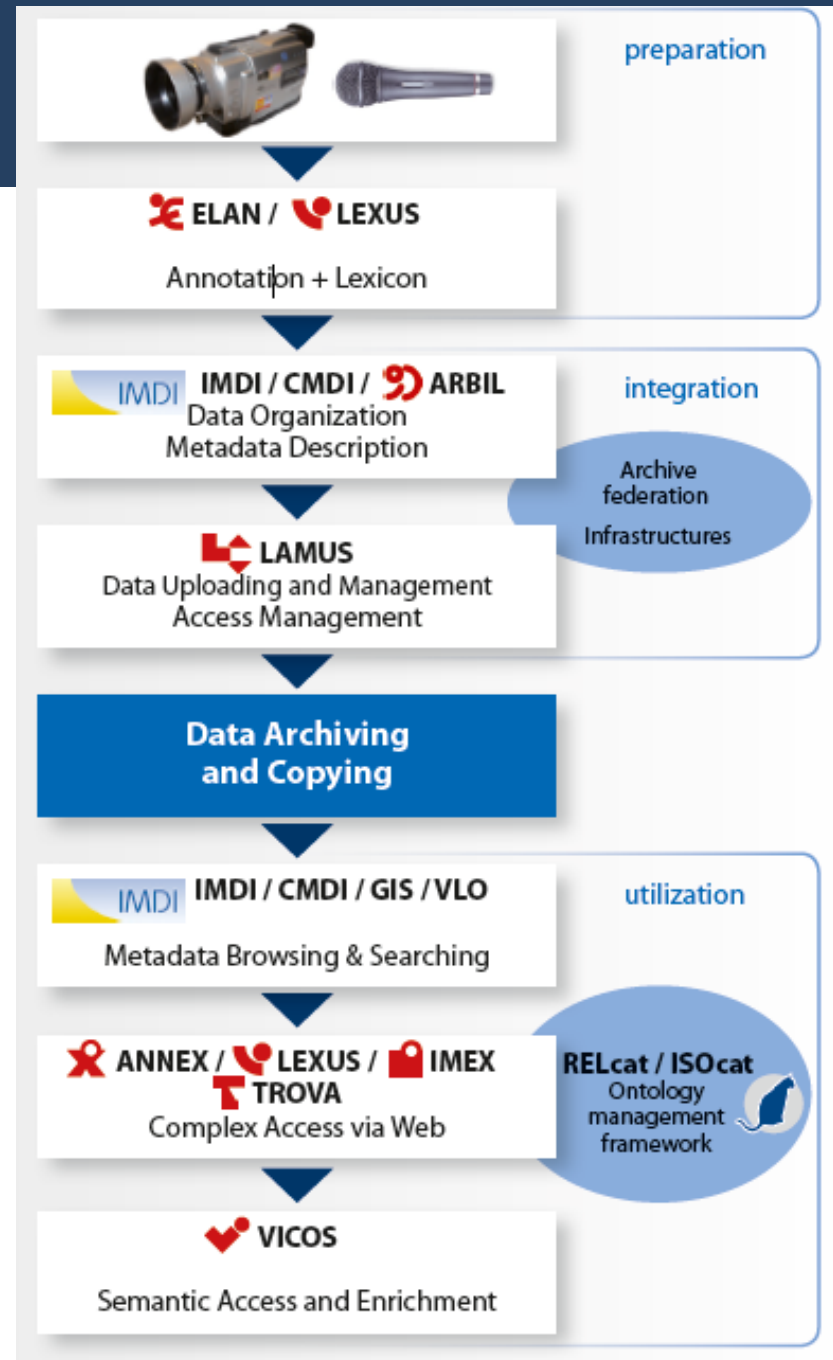


State of Archive:

- about 60 Terabyte of well-described resources
 - Linguistic & ethnographical data from 200 languages
 - Language acquisition, sign language & multimodal dt.
 - Experimental, brain imaging, genetic data etc.
- about 20.000 hours of audio/video recordings
- about 73.000 metadata described “sessions”
- about 4.5 million annotated segments
- about 50 lexica
- DOBES: ca. 25TB on 65 languages



- Full lifecycle support
- New metadata format: CMDI (flexible, components can be chosen)
- ARBIL supports CMDI
- AVATech: modules for automatic A/V recognition





- 1) Language diversity and endangerment
- 2) Language Documentation
- 3) The DOBES Programme
- 4) The Language Archive
- 5) **Relation to the Digital Humanities**



5) Relation to the Digital Humanities



- Diversity: languages, formats, approaches
- Sustainability: long-term availability, infrastructures (CLARIN etc.)
- Interdisciplinarity and teamwork
- Global network (archives, metadata)
- Open access, but also privacy rights and IPR
- Mobilization and exploitation
- DH methods for reducing annotation costs

Intl' DH Congress 2012
2012-07-19
Hamburg



Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Sebastian Drude, Paul Trilsbeek, Daan Broeder

The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands