# CMDI
# use in the
# NaLiDa project

Thorsten Trippel
thorsten.trippel@uni-tuebingen.de

Nachhaltigkeit Linguistischer Daten

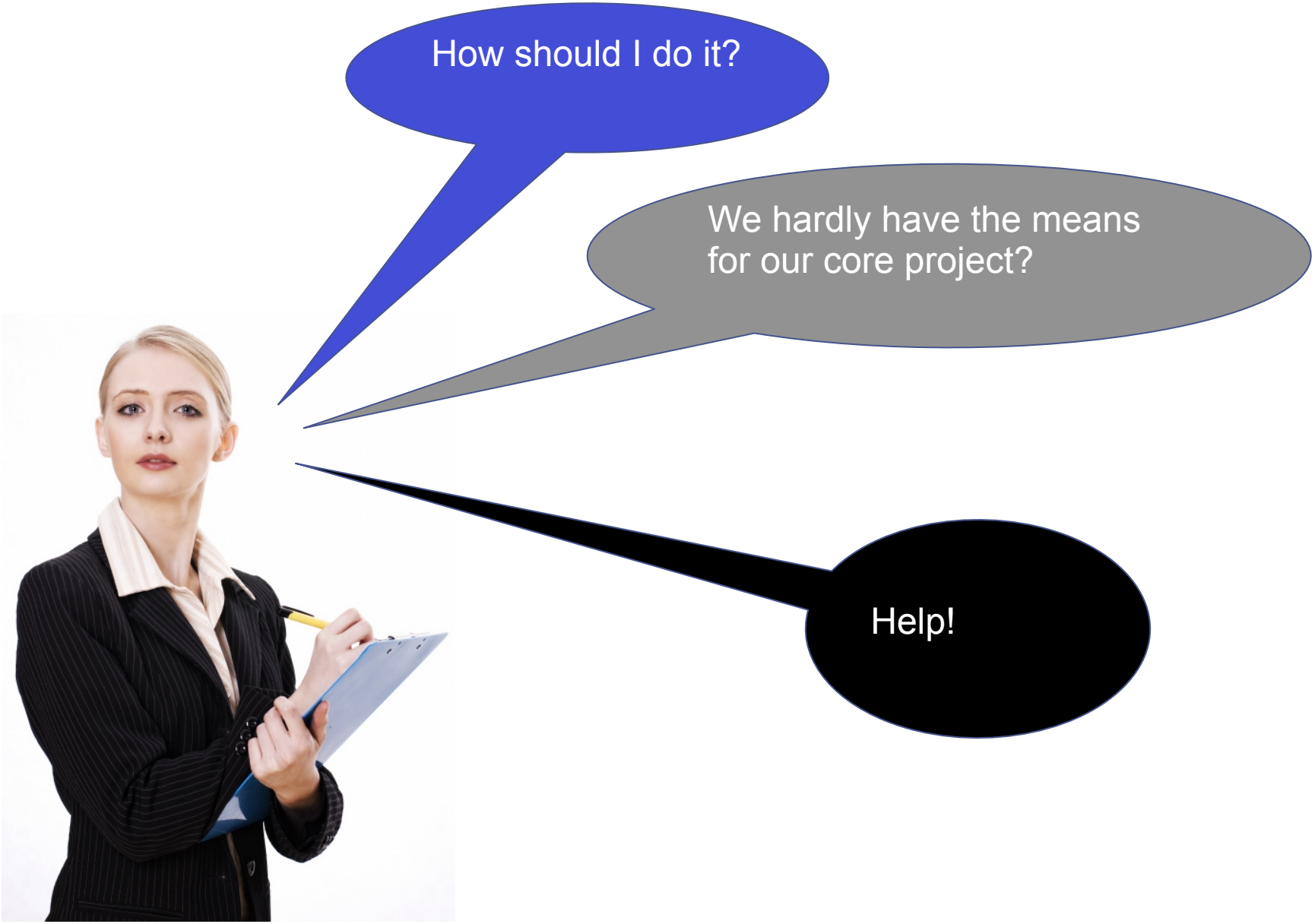
sfs
seminar für sprachwissenschaft

Eberhard Karls
Universität
Tübingen

**Archiving? Someday, but we are only at the beginning**

Nobody could use my data, it is too specialized.

I‘ll save it to 3.5“ floppy discs/CD-ROM/Exabyte ...


How should I do it?
We hardly have the means for our core project?
Help!

Worked examples

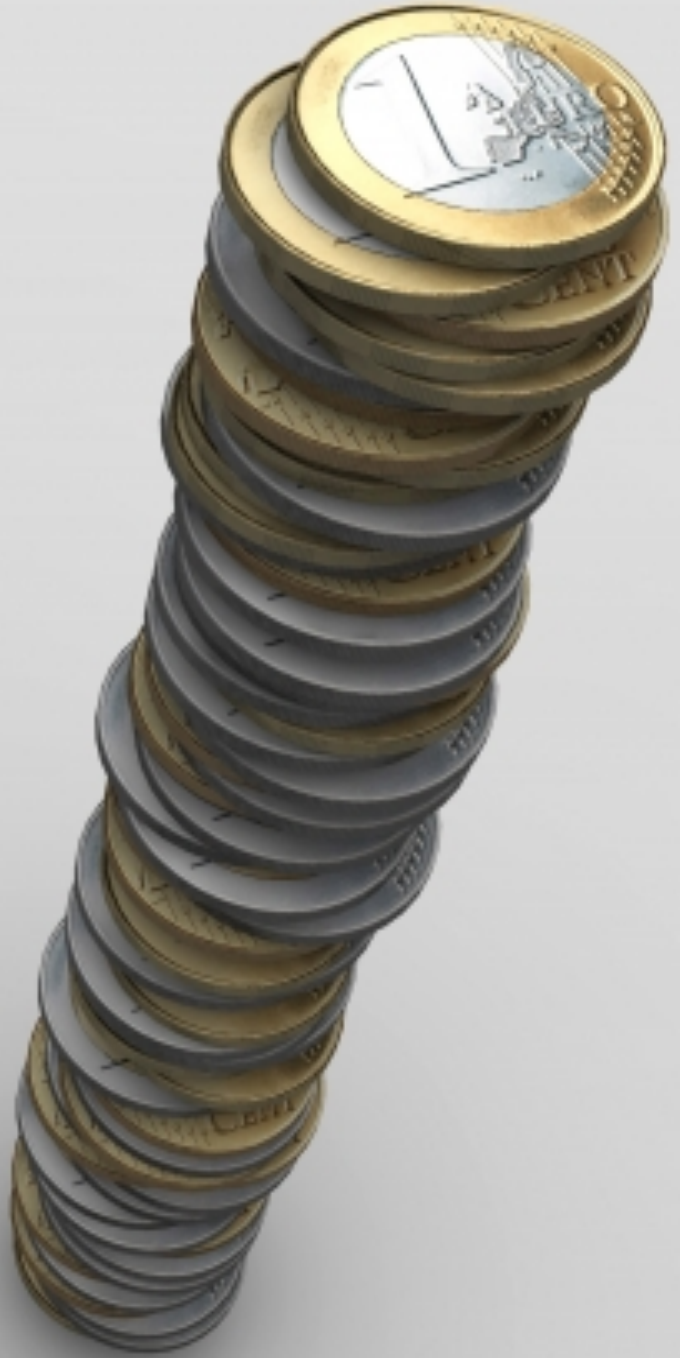
Coordination


Consulting for projects

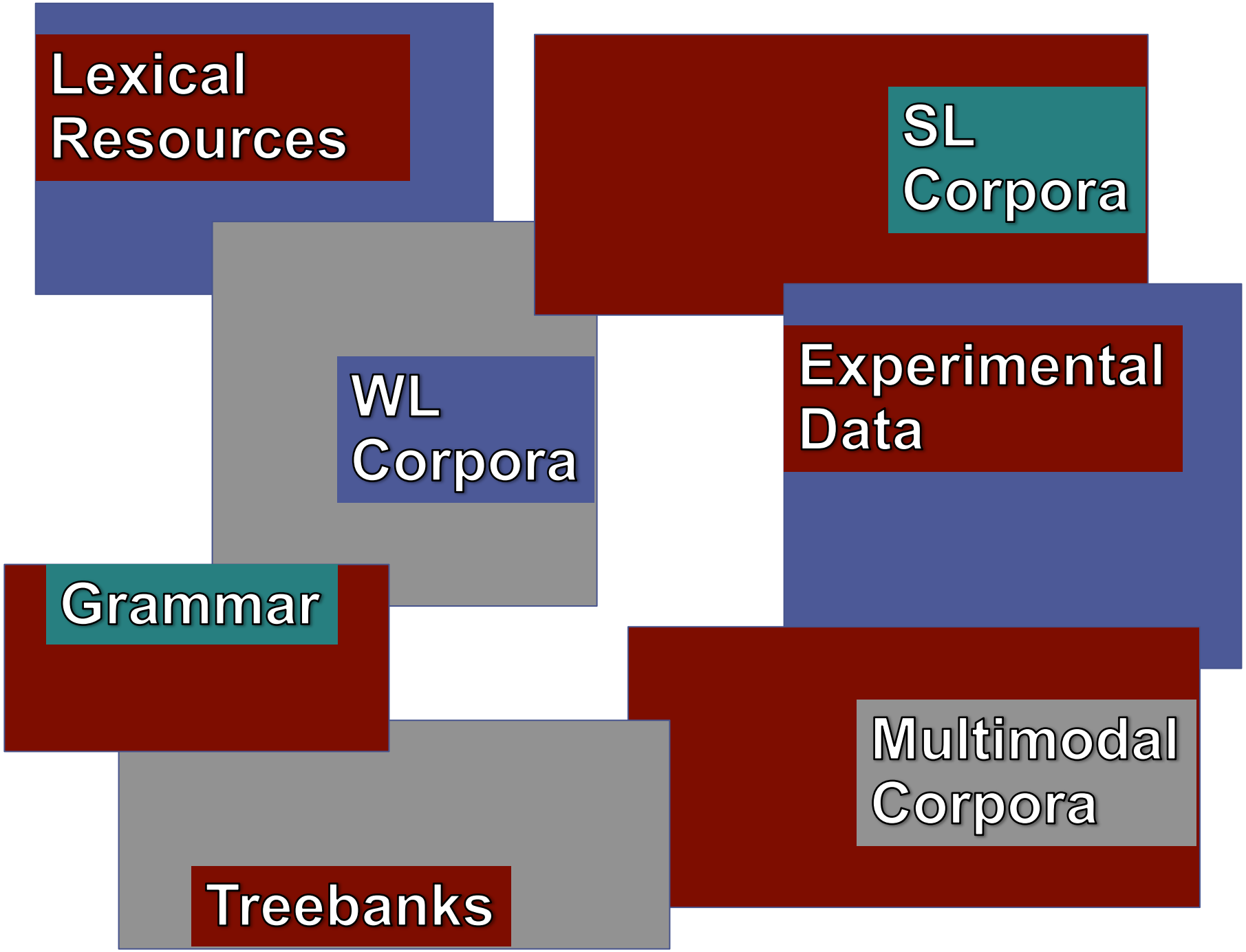
Workshops

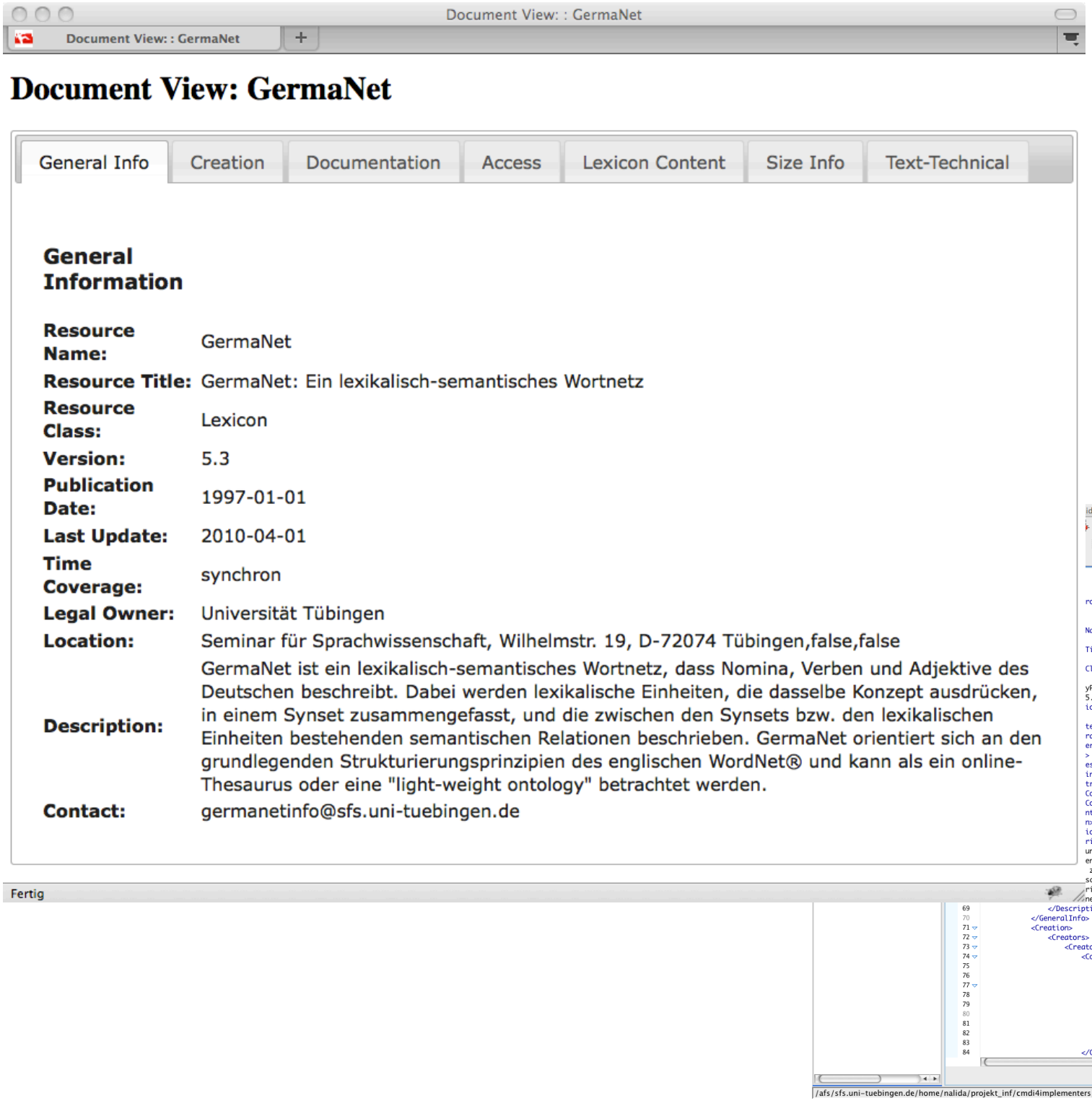
Hands on Training

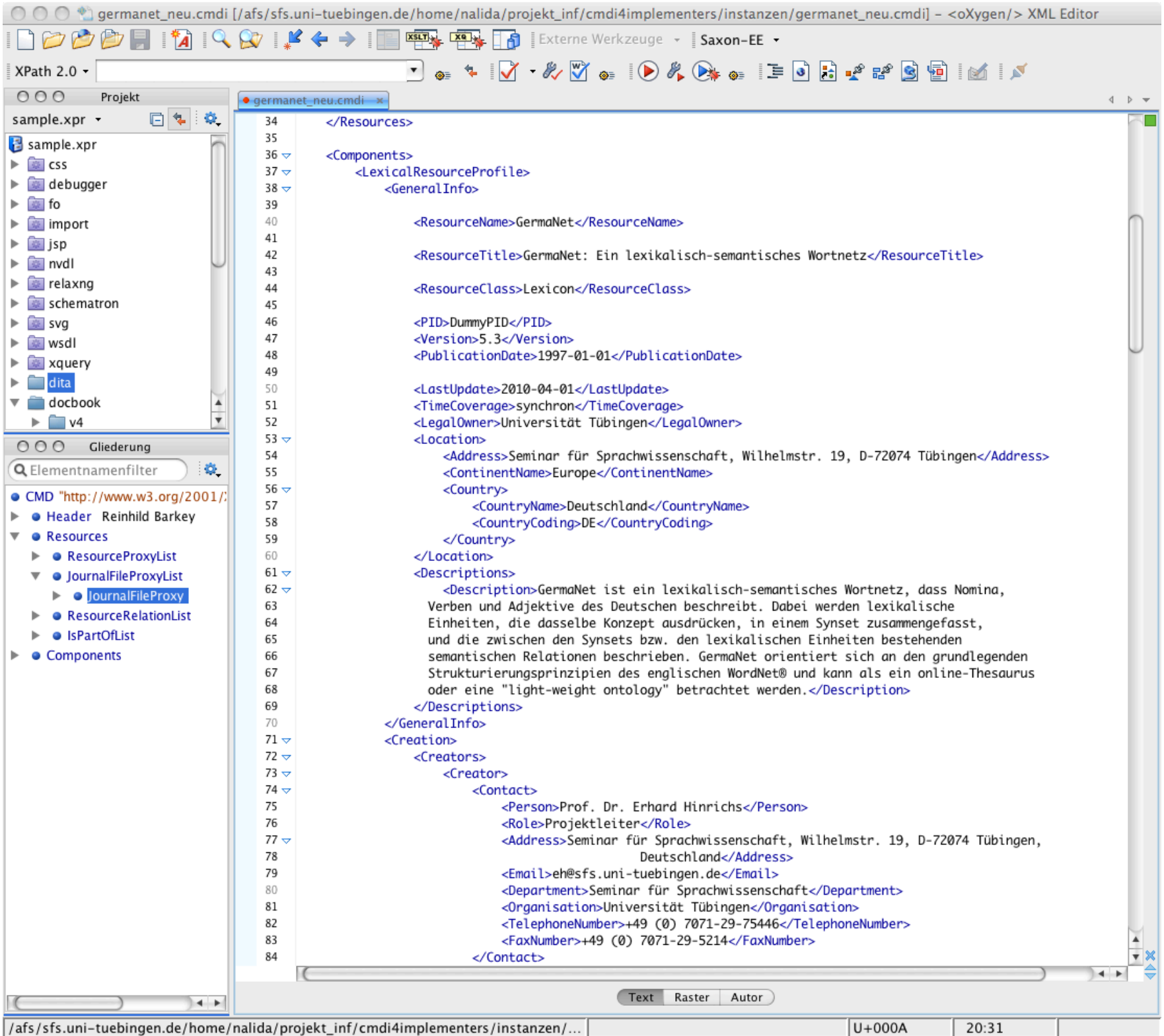
Evaluation

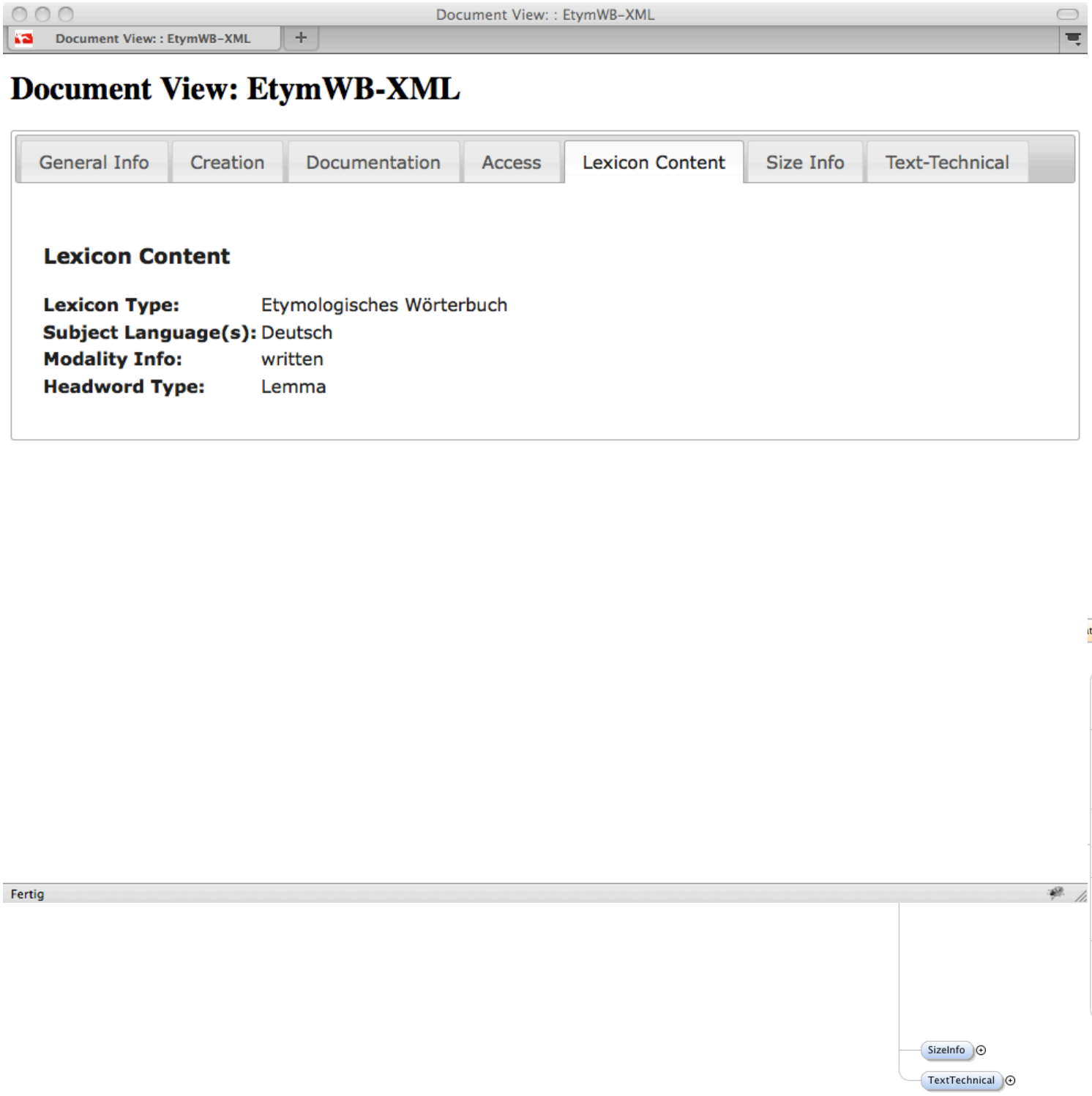


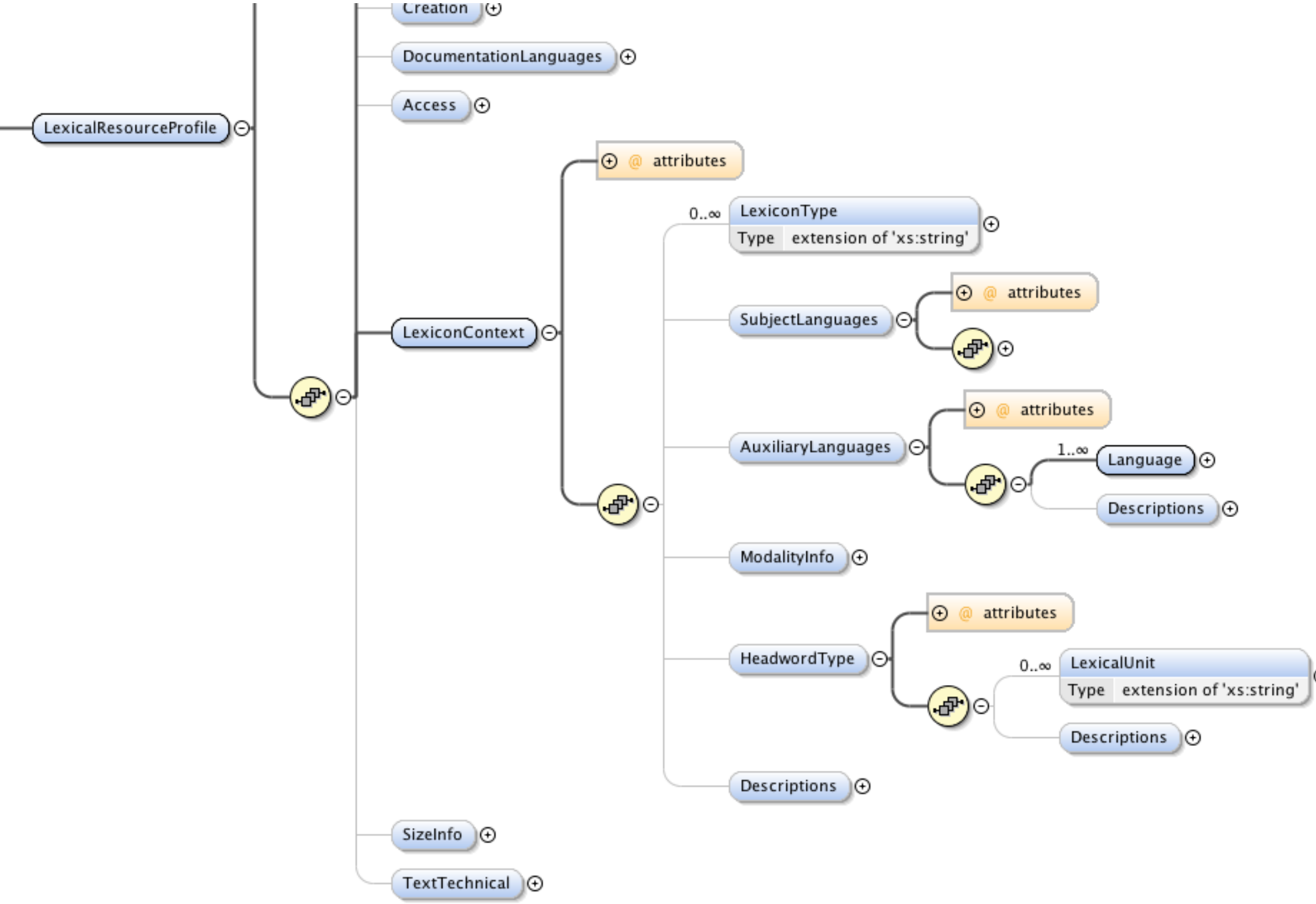

Language resources corpus german syntax NP - Google Search
Language resources corpus germ...
Web Images Videos Maps Shopping Books Mail more ▾
Web History | Search settings | Sign in
Google
Language resources corpus german syntax NP
Search
About 1,430,000 results (0.23 seconds)
Advanced search
Everything
Images
Videos
News
Shopping
More
Stuttgart
Change location
Show search tools
Corpus linguistics beyond the word: corpus research from phrase to ... - Google Books Result
Eileen Fitzpatrick - 2007 - Computers - 277 pages
It is available online via the Website of the Institute for German Language (IDS), Mannheim, Germany, with a system called COSMAS (the Corpus Search, ...
books.google.de/books?isbn=9042021357...
Statistical NLP / corpus-based computational linguistics resources
17 Feb 2010 ... Includes a POS tagger, but also NP chunking and general chunking models. .... European Language Resources Association and its catalogue. .... Saarland University Syntactically Annotated Corpus of German Newspaper Texts. ... Available (LDC), Based on Xinhua news articles. 1980s-style GB syntax. ...
www-nlp.stanford.edu/links/statnlp.html - Cached
German language - Wikipedia, the free encyclopedia
Many colonies, though, continue with German Grammar School, ..... and word groups (based on the analysis of 35 million sentences of a corpus in Leipzig, ...
en.wikipedia.org/wiki/German_language - Cached - Similar
Recent publications
In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC), pp. ... Representing SFL-annotated corpus resources. In Proceedings of the 1st .... German Grammar Fragment. Implementation of Theme for German in the ...
www.linglit.tu-darmstadt.de › ... › Personen › Teich, Elke › Forschung - Cached
Evaluating a German Sketch Grammar: A Case Study - LREC 2008 ...
by K Ivanova - Cited by 7 - Related articles
Besides the corpus itself, word sketches require a sketch grammar, a regular ... The paper presents a sketch grammar for German, a language which is not ... The evaluation focuses on NP case as a crucial part of the German grammar. ... booktitle = {Proceedings of the Sixth
Fertig

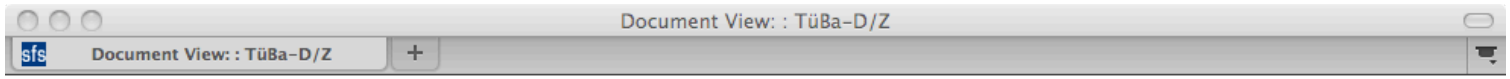

Lexical Resources
SL Corpora
WL Corpora
Experimental Data
Grammar
Multimodal Corpora
Treebanks


Document View: : GermaNet
Document View: : GermaNet
Document View: GermaNet
General Info
Creation
Documentation
Access
Lexicon Content
Size Info
Text-Technical
General Information
Resource Name: GermaNet
Resource Title: GermaNet: Ein lexikalisch-semantisches Wortnetz
Resource Class: Lexicon
Version: 5.3
Publication Date: 1997-01-01
Last Update: 2010-04-01
Time Coverage: synchron
Legal Owner: Universität Tübingen
Location: Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen,false,false
Description: GermaNet ist ein lexikalisch-semantisches Wortnetz, dass Nomina, Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische Einheiten, die dasselbe Konzept ausdrücken, in einem Synset zusammengefasst, und die zwischen den Synsets bzw. den lexikalischen Einheiten bestehenden semantischen Relationen beschrieben. GermaNet orientiert sich an den grundlegenden Strukturierungsprinzipien des englischen WordNet® und kann als ein online-Thesaurus oder eine "light-weight ontology" betrachtet werden.
Contact: germanetinfo@sfs.uni-tuebingen.de
Fertig
Saxon-EE
Text
Raster
Autor


germanet_neu.cmdi [/afs/sfs.uni-tuebingen.de/home/nalida/projekt_inf/cmdi4implementers/instanzen/germanet_neu.cmdi] – <oXygen/> XML Editor
Externe Werkzeuge
Saxon-EE
XPath 2.0
Projekt
sample.xpr
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
v4
Gliederung
Elementnamenfilter
CMD "http://www.w3.org/2001/
Header Reinhild Barkey
Resources
ResourceProxyList
JournalFileProxyList
JournalFileProxy
ResourceRelationList
IsPartOfList
Components
germanet_neu.cmdi
34 </Resources>
35
36 <Components>
37 <LexicalResourceProfile>
38 <GeneralInfo>
39
40 <ResourceName>GermaNet</ResourceName>
41
42 <ResourceTitle>GermaNet: Ein lexikalisch-semantisches Wortnetz</ResourceTitle>
43
44 <ResourceClass>Lexicon</ResourceClass>
45
46 <PID>DummyPID</PID>
47 <Version>5.3</Version>
48 <PublicationDate>1997-01-01</PublicationDate>
49
50 <LastUpdate>2010-04-01</LastUpdate>
51 <TimeCoverage>synchron</TimeCoverage>
52 <LegalOwner>Universität Tübingen</LegalOwner>
53 <Location>
54 <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen</Address>
55 <ContinentName>Europe</ContinentName>
56 <Country>
57 <CountryName>Deutschland</CountryName>
58 <CountryCoding>DE</CountryCoding>
59 </Country>
60 </Location>
61 <Descriptions>
62 <Description>GermaNet ist ein lexikalisch-semantisches Wortnetz, dass Nomina,
63 Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische
64 Einheiten, die dasselbe Konzept ausdrücken, in einem Synset zusammengefasst,
65 und die zwischen den Synsets bzw. den lexikalischen Einheiten bestehenden
66 semantischen Relationen beschrieben. GermaNet orientiert sich an den grundlegenden
67 Strukturierungsprinzipien des englischen WordNet® und kann als ein online-Thesaurus
68 oder eine "light-weight ontology" betrachtet werden.</Description>
69 </Descriptions>
70 </GeneralInfo>
71 <Creation>
72 <Creators>
73 <Creator>
74 <Contact>
75 <Person>Prof. Dr. Erhard Hinrichs</Person>
76 <Role>Projektleiter</Role>
77 <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen,
78 Deutschland</Address>
79 <Email>eh@sfs.uni-tuebingen.de</Email>
80 <Department>Seminar für Sprachwissenschaft</Department>
81 <Organisation>Universität Tübingen</Organisation>
82 <TelephoneNumber>+49 (0) 7071-29-75446</TelephoneNumber>
83 <FaxNumber>+49 (0) 7071-29-5214</FaxNumber>
84 </Contact>
Text
Raster
Autor
/afs/sfs.uni-tuebingen.de/home/nalida/projekt_inf/cmdi4implementers/instanzen/...
U+000A
20:31


Document View: : EtymWB-XML
Document View: : EtymWB-XML
Document View: EtymWB-XML
General Info
Creation
Documentation
Access
Lexicon Content
Size Info
Text-Technical
Lexicon Content
Lexicon Type: Etymologisches Wörterbuch
Subject Language(s): Deutsch
Modality Info: written
Headword Type: Lemma
Fertig
ttributes
0..∞
LexiconType
Type extension of 'xs:string'
SubjectLanguages
@ attributes
AuxiliaryLanguages
@ attributes
1..∞
Language
Descriptions
ModalityInfo
HeadwordType
@ attributes
0..∞
LexicalUnit
Type extension of 'xs:string'
Descriptions
Descriptions
SizeInfo
TextTechnical


Creation
DocumentationLanguages
Access
LexicalResourceProfile
@ attributes
0..∞
LexiconType
Type extension of 'xs:string'
LexiconContext
@ attributes
SubjectLanguages
@ attributes
AuxiliaryLanguages
1..∞
Language
Descriptions
ModalityInfo
@ attributes
HeadwordType
0..∞
LexicalUnit
Type extension of 'xs:string'
Descriptions
Descriptions
SizeInfo
TextTechnical




# Document View: TüBa-D/Z

General Info | Project | CorpusContext | Creation | Access | Subject Languages | Documentation Languages | Tagset Info | Size Info | Modality Info | Validation Group | Text-Technical

## Access

| | |
|---|---|
| **Availability:** | Registrierung erforderlich, die Texte der taz müssen erworben werden |
| **Distribution Medium:** | Download |
| **Catalogue Link:** | http://www.sfs.uni-tuebingen.de/tuebadz.shtml |
| **Price:** | free for academic use, registration required; restricted use for non-academic R&D and commercial use; requires taz-archive licence |
| **Deployment Tool Info:** | TIGERSearch |
| **Contact:** | Kathrin Beck |
| **Descriptions:** | TIGERSearch ist ein Programm zur Untersuchung syntaktisch annotierter Textkorpora. Korpora, die als Syntaxbäume dargestellt werden, können über eine graphische Oberfläche nach bestimmten Baumstrukturen durchsucht werden. TIGERSearch ist speziell zur Suche auf der deutschsprachigen TIGER-Baumbank gedacht, kann aber auch auf andere Baumbanken angewendet werden. |

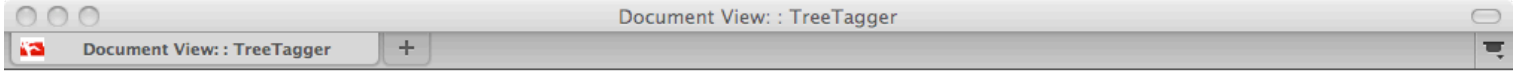
str. 19, D-72074 Tübingen</Address>

ank des Deutschen / Schriftsprache (TüBa-D/Z) ist ein syntaktisch
eitung "die tageszeitung" (taz).
unabhängig.
benen syntaktischer Konstituenz: die lexikalische
opologischen Felder und die Satzebene.

Fertig

```
66              <CorpusType>general corpus</CorpusType>
67              <TemporalClassification>synchronic</TemporalClassification>
68          </CorpusContext>
69              <Creation>
70                  <OriginalSource>taz DVD</OriginalSource>
71                  <SourceType>Tageszeitung</SourceType>
72                  <Creators>
73                      <Creator>
74                        <Contact>
75                          <Person>Prof. Dr. Erhard Hinrichs</Person>
76                          <Role>Projektleiter</Role>
77                          <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen, Deutschland</Address>
78                          <Email>eh@sfs.uni-tuebingen.de</Email>
79                          <Department xml:lang="de">Seminar für Sprachwissenschaft (SfS)</Department>
80                          <Organisation xml:lang="de">Universität Tübingen</Organisation>
81                          <TelephoneNumber>+49 (0) 7071-29-75446</TelephoneNumber>
82                          <FaxNumber>+49 (0) 7071-29-5214</FaxNumber>
```

```
                <PID></PID>
                <Version>5</Version>
                <StartYear>1999</StartYear>
                <PublicationDate>November 2009</PublicationDate>
                <LastUpdate>2009-11-30</LastUpdate>
                <LegalOwner>Universität Tübingen</LegalOwner>
                <Location>
                    <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen</Address>
                    <ContinentName>Europe</ContinentName>
                     <Country>
                        <CountryName>Deutschland</CountryName>
                        <CountryCoding>DE</CountryCoding>
                     </Country>
                 </Location>
                 <Descriptions>
                       <Description xml:lang="de">Die Tübinger Baumbank des Deutschen / Schriftsprache (TüBa-D/Z) ist ein syntaktisch
                          annotiertes Korpus auf der Grundlage der Zeitung "die tageszeitung" (taz).
                          Die Annotation ist (weitestgehend) theorieunabhängig.
                          Das Annotationsschema unterscheidet vier Ebenen syntaktischer Konstituenz: die lexikalische
                          Ebene, die phrasale Ebene, die Ebene der topologischen Felder und die Satzebene.
                       </Description>
                </Descriptions>
            </GeneralInfo>
        <CorpusContext>
            <CorpusType>general corpus</CorpusType>
            <TemporalClassification>synchronic</TemporalClassification>
        </CorpusContext>
            <Creation>
                <OriginalSource>taz DVD</OriginalSource>
                <SourceType>Tageszeitung</SourceType>
                <Creators>
                     <Creator>
                       <Contact>
                         <Person>Prof. Dr. Erhard Hinrichs</Person>
                         <Role>Projektleiter</Role>
                         <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen, Deutschland</Address>
                         <Email>eh@sfs.uni-tuebingen.de</Email>
                         <Department xml:lang="de">Seminar für Sprachwissenschaft (SfS)</Department>
                         <Organisation xml:lang="de">Universität Tübingen</Organisation>
                         <TelephoneNumber>+49 (0) 7071-29-75446</TelephoneNumber>
                         <FaxNumber>+49 (0) 7071-29-5214</FaxNumber>
```



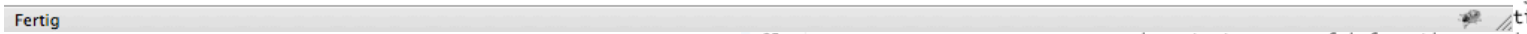

# Document View: TreeTagger

General Info | Project | Creators | Access | Copyright | Resource Context | Tool Context

## Project

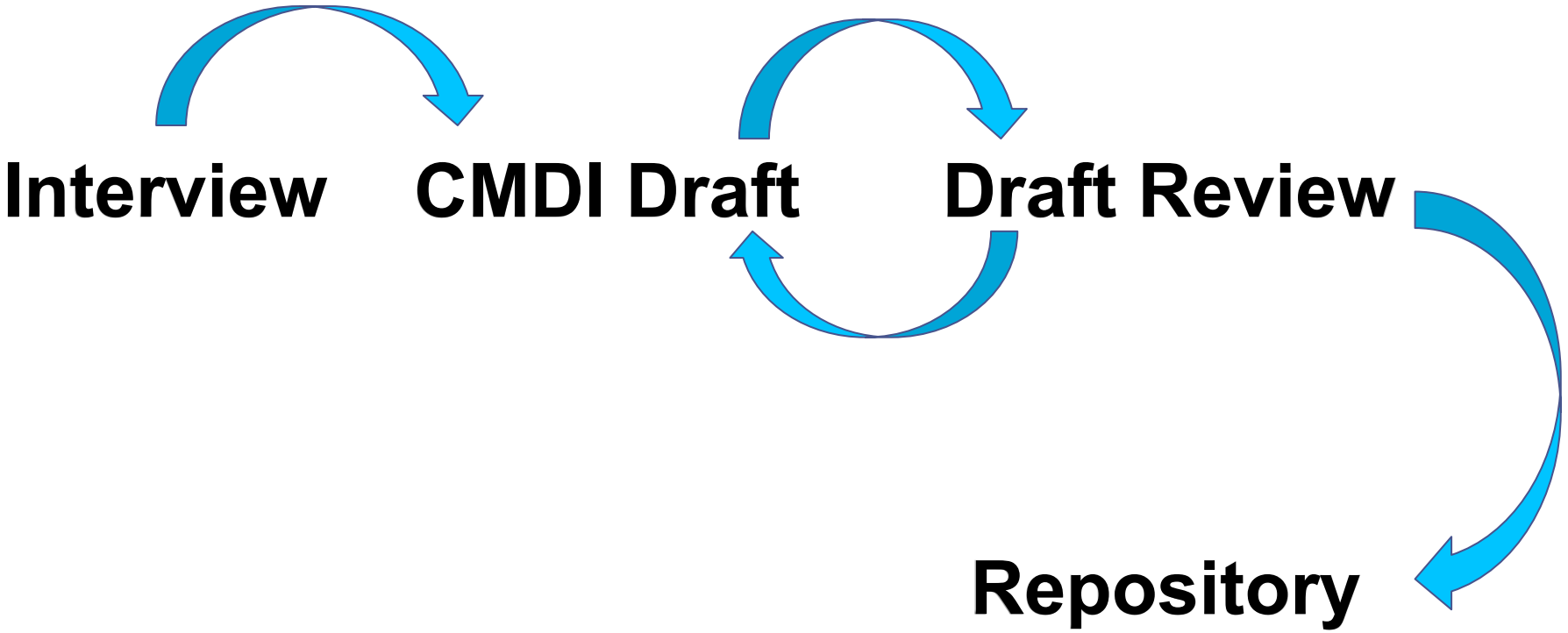
| | |
|---|---|
| **Project Name:** | TC Project |
| **Project Title:** | false |
| **Project ID:** | false |
| **Funder:** | Ministry of Science and Research of the Land Baden-Württemberg |
| **Url:** | http://www.ims.uni-stuttgart.de/projekte/tc/ |
| **Institution:** | false,false,Germany |
| **Cooperations:** | false |
| **Descriptions:** | In a world of growing floods of information, humans need the assistance of mechanized information processing tools. One important source of information consists of written or spoken text. The Institute for Natural Language Processing (IMS) carries out basic and applied research and trains students to create tools for automated processing of spoken and written language. |
| **Contact:** | Helmut.Schmid@ims.uni-stuttgart.de |
| **Duration:** | 1993-1996 |

```
Tools for Their Exploration</ProjectTitle>
hließungswerkzeuge</ProjectTitle>

Land Baden-Württemberg</Funder>
/</Url>

inelle Sprachverarbeitung (IMS)</Department>
ral Language Processing (IMS)</Department>
tgart</Organisation>
uttgart</Organisation>
art.de/ims-home.html.en</Url>
art.de/ims-home.html.de</Url>

tgart</Address>
```




```
growing floods of information, humans need
tion processing tools. One
95                important source of information consists of written or spoken text.
96                The Institute for Natural Language Processing (IMS) carries out
97                basic and applied research and trains students to create tools for
98                automated processing of spoken and written language.</Description>
99            </Descriptions>
100           <Contact>
101               <Person>Sabine Dieterle</Person>
102               <Role>Secretary</Role>
103               <Address>Azenbergstraße 12, D-70174 Stuttgart, Germany</Address>
104               <Email>ims@ims.uni-stuttgart.de</Email>
105               <Department xml:lang="de">Institut für Maschinelle Sprachverarbeitung (IMS)</Department>
106               <Department xml:lang="en">Institute for Natural Language Processing (IMS)</Department>
107               <Organisation xml:lang="de">Universität Stuttgart</Organisation>
108               <Organisation xml:lang="en">University of Stuttgart</Organisation>
109               <TelephoneNumber>+49 711 685 81379</TelephoneNumber>
```

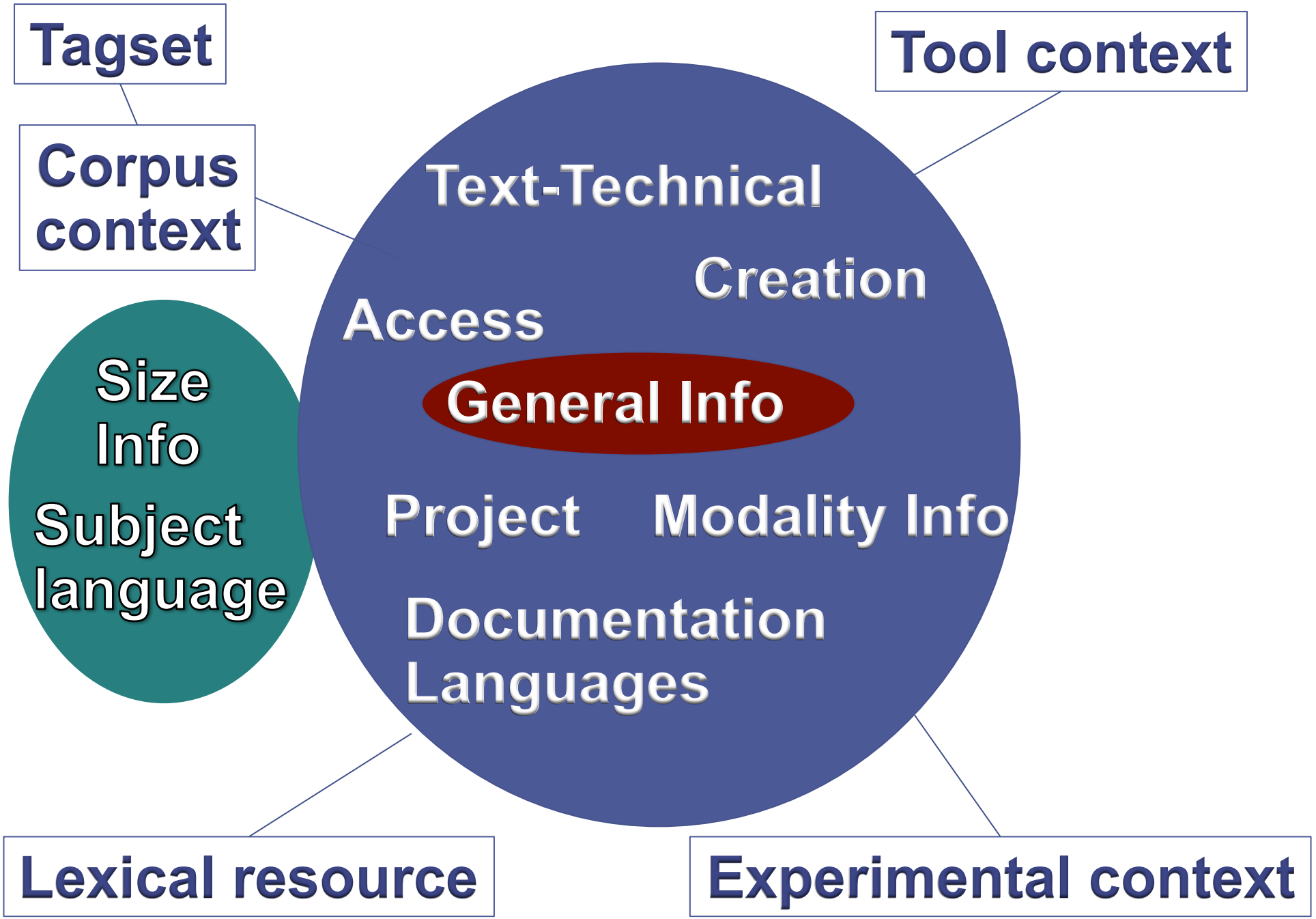
```
<Project>
    <ProjectName>TC Project</ProjectName>
    <ProjectTitle xml:lang="en">Textual Corpora and Tools for Their Exploration</ProjectTitle>
    <ProjectTitle xml:lang="de">Textkorpora und Erschließungswerkzeuge</ProjectTitle>
    <ProjectID></ProjectID>
    <Funder>Ministry of Science and Research of the Land Baden-Württemberg</Funder>
    <Url>http://www.ims.uni-stuttgart.de/projekte/tc/</Url>
    <Institution>
        <Department xml:lang="de">Institut für Maschinelle Sprachverarbeitung (IMS)</Department>
        <Department xml:lang="en">Institute for Natural Language Processing (IMS)</Department>
        <Organisation xml:lang="de">Universität Stuttgart</Organisation>
        <Organisation xml:lang="en">University of Stuttgart</Organisation>
        <Url xml:lang="en">http://www.ims.uni-stuttgart.de/ims-home.html.en</Url>
        <Url xml:lang="de">http://www.ims.uni-stuttgart.de/ims-home.html.de</Url>
        <Location>
            <Address>Azenbergstraße 12, D-70174 Stuttgart</Address>
            <Region></Region>
            <ContinentName>Europe</ContinentName>
            <Country>
                <CountryName>Germany</CountryName>
                <CountryCoding>DE</CountryCoding>
            </Country>
        </Location>
        <Descriptions>
            <Description xml:lang="en">In a world of growing floods of information, humans need
                the assistance of mechanized information processing tools. One
                important source of information consists of written or spoken text.
                The Institute for Natural Language Processing (IMS) carries out
                basic and applied research and trains students to create tools for
                automated processing of spoken and written language.</Description>
        </Descriptions>
        <Contact>
            <Person>Sabine Dieterle</Person>
            <Role>Secretary</Role>
            <Address>Azenbergstraße 12, D-70174 Stuttgart, Germany</Address>
            <Email>ims@ims.uni-stuttgart.de</Email>
            <Department xml:lang="de">Institut für Maschinelle Sprachverarbeitung (IMS)</Department>
            <Department xml:lang="en">Institute for Natural Language Processing (IMS)</Department>
            <Organisation xml:lang="de">Universität Stuttgart</Organisation>
            <Organisation xml:lang="en">University of Stuttgart</Organisation>
            <TelephoneNumber>+49 711 685 81379</TelephoneNumber>
```

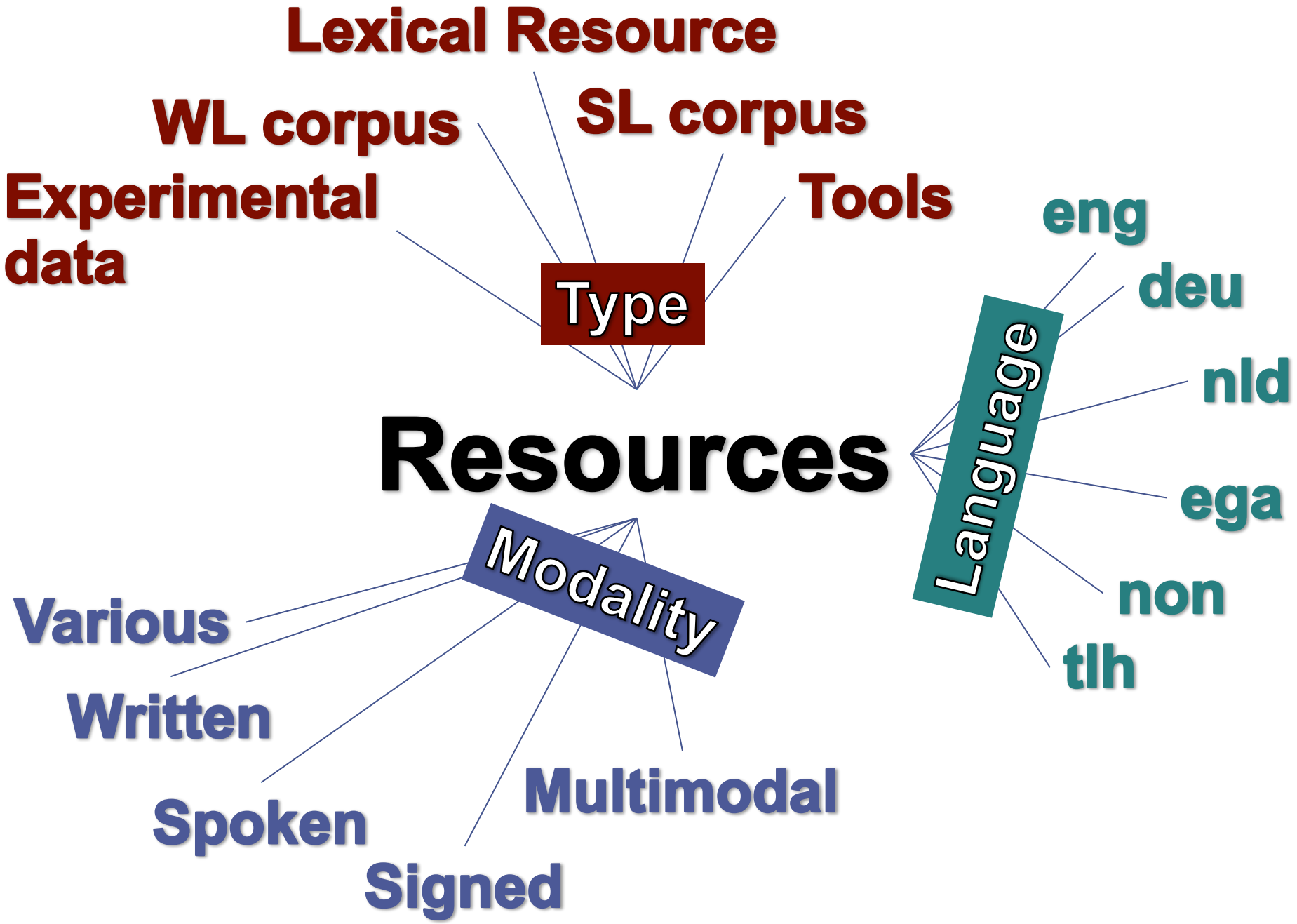

Interview
CMDI Draft
Draft Review
Repository

# CMDI Data in NaLiDa

| Type | Manual | Transformed | Harvested |
|---|---|---|---|
| Lexical resources | 2 | 6 | 0** |
| Corpora | 2 | 55 | ~2800* |
| Tools | 2 | 0* | 145* |
| Webservices | 2 | 119 | 0* |
| Experimental Data | 0* | 0 | 0 |
| | | | >10000 |

*: not originally in CMDI
**: classification in harvested data not available, often counted among corpora


Tagset
Tool context
Corpus context
Text-Technical
Creation
Access
General Info
Size Info
Subject language
Project
Modality Info
Documentation Languages
Lexical resource
Experimental context


Lexical Resource
WL corpus
SL corpus
Experimental data
Tools
Type
Resources
Language
eng
deu
nld
ega
non
tlh
Modality
Various
Written
Spoken
Signed
Multimodal

# The NaLiDa contribution

- Added value
- Findability
- Structured description
- Long term archiving
- Prototypical examples

# Credits/Photos

Nachhaltigkeit Linguistischer Daten

Petroleum Pipes Image: Suat Eman / FreeDigitalPhotos.net

Railway Image: samurai / FreeDigitalPhotos.net

Lake Vyrnwy Dam Image: Matt Banks / FreeDigitalPhotos.net

Man reading Image: Arvind Balaraman / FreeDigitalPhotos.net

Business Woman Taking Notes Image: Michal Marcol / FreeDigitalPhotos.net

Euro In Pila Image: Idea go / FreeDigitalPhotos.net