

Vagueness in Metadata

Nicole Majka^{1,2}, Gabriele Schwiertz¹ & Felix Rau¹

¹ University of Cologne

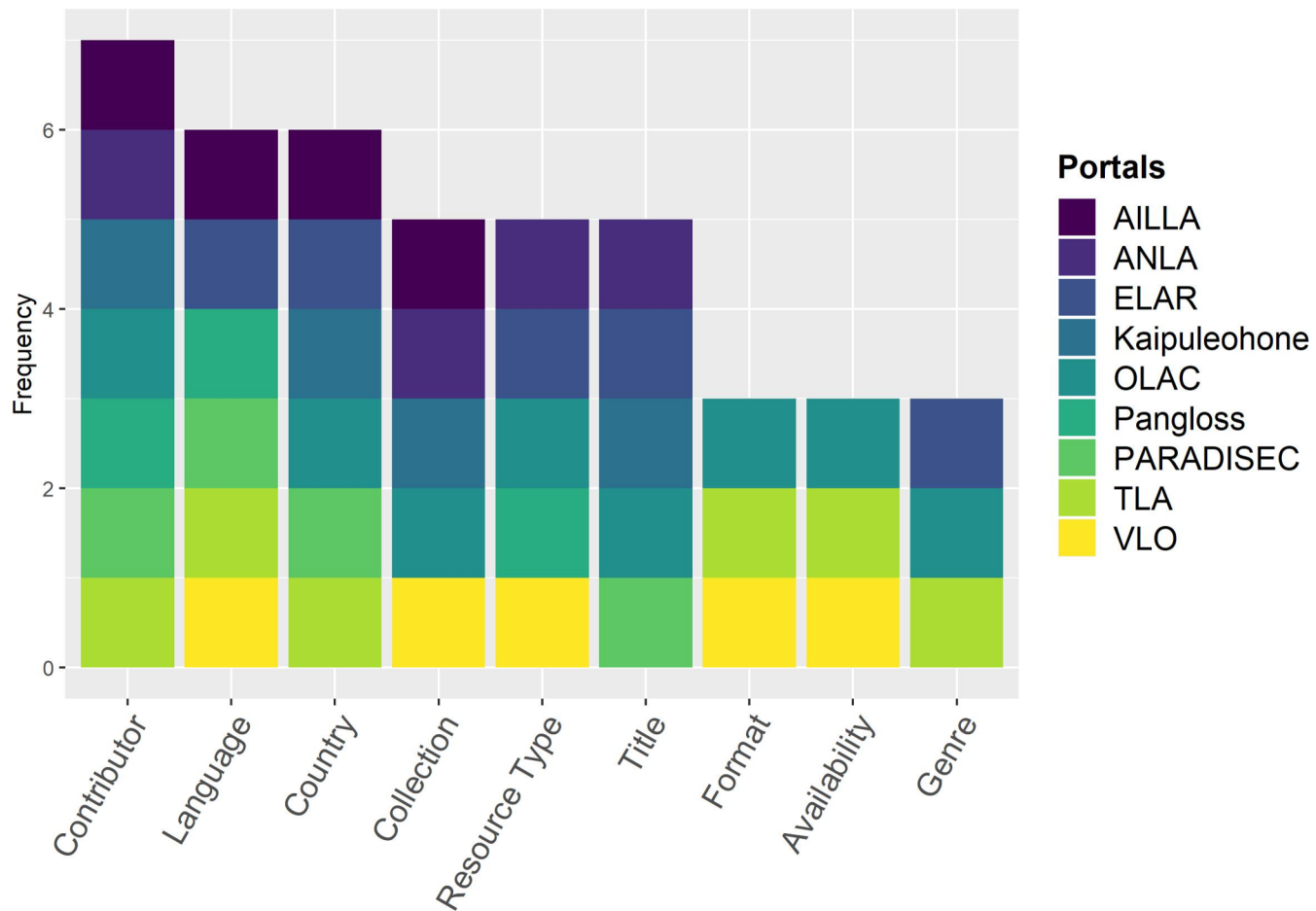
² Bangor University

Outline

1. Two surveys into metadata usage
2. Example: Language
3. Example: Location
4. What is vagueness?
5. How does vagueness impact findability and reliability?
6. Solutions

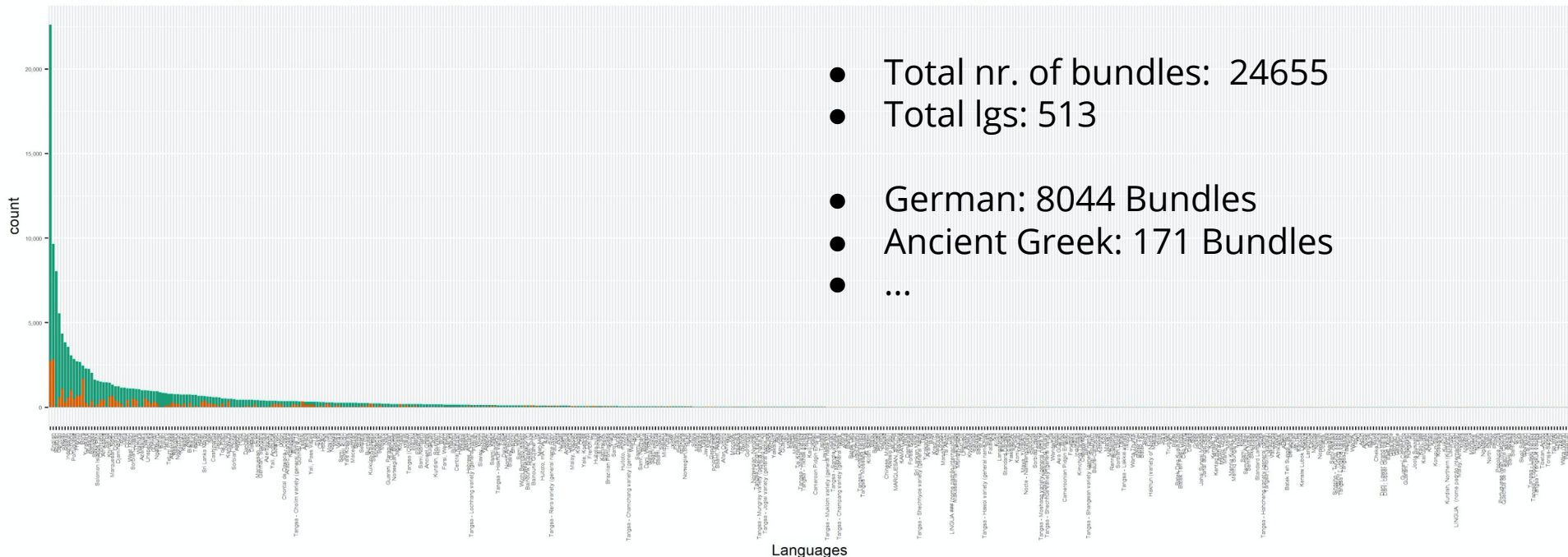
2 surveys into metadata usage

Portal Survey

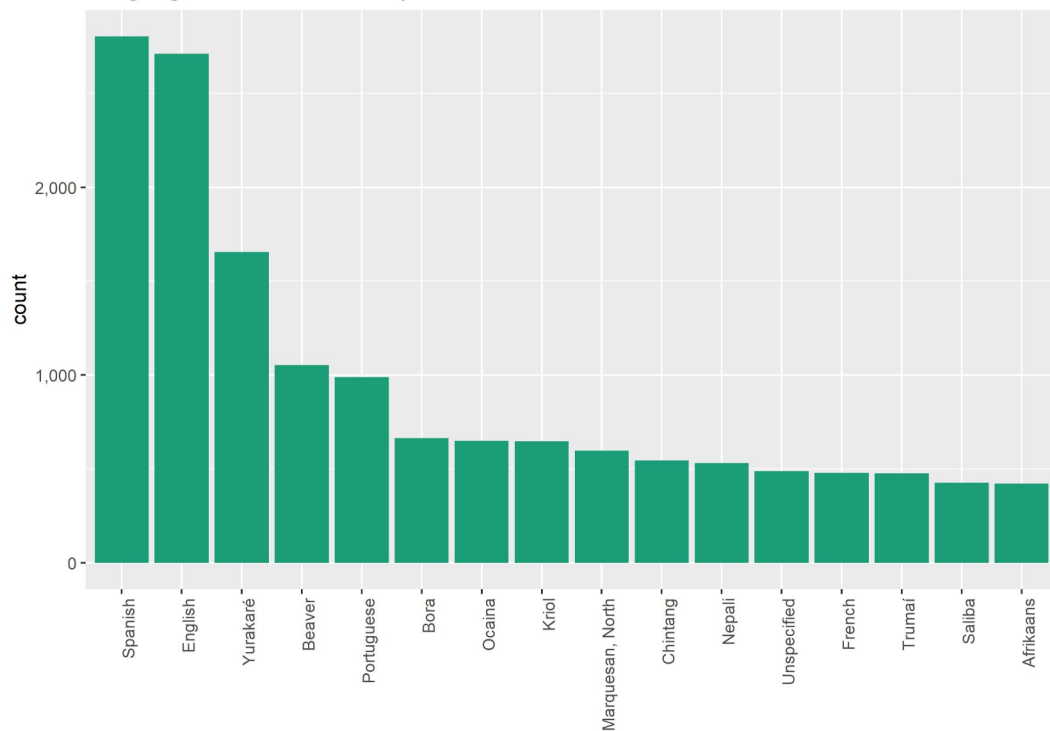


IMDI Survey: Languages

Language mentions in the DoBeS corpus



IMDI Survey: Top 10 Languages



Trying to be more specific:
ContentLanguages

- *Spanish*
- *English*
- Yurakaré
- Beaver
- *Portuguese*
- Bora
- Ocaina
- Kriol
- ...

2 examples: Language & Location

Example: Language

Semantics of Language

- Content Language: language the resource is in
- Subject Language: language the resource is about
- Actor Language: language used or known by a participant
- Translation Language: language the resource has been translated into

Example: Language

Type of Language	Schema.org	DataCite	OLAC	IMDI	IDS	HZSK	BLAM
Subject Language	(✓)	(✓)	✓	✓	✓	✓	✓
Content Language	✓	✓	✓	✓	(✓)	✓	
Actor Language	✓			✓	✓	✓	
Translation Language	✓			✓	(✓)	✓	✓

Example: Location

Semantics of Location:

- Location of the resource: Where is the resource stored?
- Recording location: Where was the resource recorded?
- Location of subject language or dialect
- Location of Actor: Where does the participant come from? Where does the participant live?

Example: Location

Type of Location	Schema.org	DataCite	OLAC	IMDI	IDS	HZSK	BLAM
Location of Resource	✓				✓		
Recording Location	✓	✓	✓	✓	✓	✓	✓
Location of Language or Dialect		✓	✓	✓	(✓)	✓	✓
Location of Actor	✓				✓	✓	

What is vagueness?

What is vagueness?

In the title, we use vagueness to cover the four types:

- vagueness
- generality
- uncertainty
- ambiguity

Vagueness and Generality

Vagueness:

1 grain of rice is not a heap, 1+1 grains, 2+1, 100000 +1, 1000000 + 1, ...

DataCite Language “The primary language of the resource.”

Generality:

“I saw a reptile.” vs. “I saw a snake.”

German (deu), Alemannic (gsw), Swiss German (gsw), Bernese (gsw)

Uncertainty and Ambiguity

Uncertainty:

"This is probably edible." (introduced mostly through *optional fields*)

Ambiguity:

"Yesterday, I went to the bank."

Kui (uki) India

Kui (kvd) Indonesia

Kui (kdt) Thailand/Laos/Cambodia

Austroasiatic

Dravidian

Timor-Alor-Pantar

How does vagueness impact findability and reliability?

Findability

DataCite (Language)

german

Search

542,539 Works

Zur Identifikation leerer Subjekte in infinitivischen Komplementsätzen – ein semantisch-pragmatisches Modell

Klaus-Michael Köpcke & Klaus-Uwe Panther
Article published via Berlin : Walter de Gruyter

The interpretation of empty elements, i.e. signs that have no phonetic realization, constitutes a "classical" problem in modern linguistics. Null elements have been postulated on various levels of the linguistic system and its pragmatic use, in particular, in morphology and syntax. An adequate theory of discourse comprehension also requires an account of what is not said but only conversationally implicated. In the last thirty years the interpretation of empty subjects in non-finite clauses, which is...

No citations were reported. No usage information was reported.

<https://doi.org/10.15488/207> Cite

Annotationsrichtlinien des Projekts "Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse"

Annelen Brunner, Lukas Weimer, Stefan Engelberg, Fotis Jannidis & Ngoc Duyen Tanja Tu
Working Paper published via Zenodo

Annotationsrichtlinien für die Wiedergabe von Rede, Gedanken und Geschriebenem, verwendet im DFG-Projekt "Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse" (redewiedergabe.de) zur Erstellung des Redewiedergabe-Korpus.

No citations were reported. No usage information was reported.

<https://doi.org/10.5281/zenodo.2634995> Cite

Registration Year

<input type="checkbox"/> 2021	1,851
<input type="checkbox"/> 2020	34,530
<input type="checkbox"/> 2019	31,397
<input type="checkbox"/> 2018	22,366
<input type="checkbox"/> 2017	16,285
<input type="checkbox"/> 2016	12,299
<input type="checkbox"/> 2015	11,786
<input type="checkbox"/> 2014	265,044
<input type="checkbox"/> 2013	123,527
<input type="checkbox"/> 2012	6,308

Resource Types

<input type="checkbox"/> Text	474,271
<input type="checkbox"/> Dataset	54,317
<input type="checkbox"/> Collection	4,094
<input type="checkbox"/> Other	2,281
<input type="checkbox"/> Audiovisual	1,575
<input type="checkbox"/> Image	1,047
<input type="checkbox"/> Software	436
<input type="checkbox"/> Workflow	295
<input type="checkbox"/> Interactive Resource	112
<input type="checkbox"/> Sound	30
<input type="checkbox"/> Physical Object	27
<input type="checkbox"/> Event	24
<input type="checkbox"/> Data Paper	18
<input type="checkbox"/> Model	6


Findability

DataCite
(Language)

Обучение деловой переписке на немецком языке специалистов с дополнительной квалификацией «Переводчик в сфере профессиональной коммуникации»

Journal Article published via Научно-методический электронный журнал «Концепт»

В статье рассматриваются актуальные проблемы обучения деловой переписке на немецком языке специалистов с дополнительной квалификацией «Переводчик в сфере профессиональной коммуникации». Выделен комплекс основных тематических аспектов, подлежащих обучению, объясняются цели и задачи авторского курса, приводятся рекомендации по соблюдению речевых, стилистических и грамматических особенностей написания делового письма на немецком языке согласно правилам, установленным Немецким институтом стандартизации (DIN), а также примеры стилистических различий при составлении писем на немецком и русском языках. Раскрываются приемы и методы работы, приводятся примеры...

 No citations were reported. No usage information was reported.

 <https://doi.org/10.24422/mcito.2017.v8.6986>  Cite

Findability – DataCite (Language)

Registration Year		Resource Types		Affiliations	
<input type="checkbox"/> 2021	1,851	<input type="checkbox"/> Text	474,271	<input type="checkbox"/> Ludwig-Maximilians-Universität München	202
<input type="checkbox"/> 2020	34,530	<input type="checkbox"/> Dataset	54,317	<input type="checkbox"/> Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences	21
<input type="checkbox"/> 2019	31,397	<input type="checkbox"/> Collection	4,094	<input type="checkbox"/> TU Wien	19
<input type="checkbox"/> 2018	22,366	<input type="checkbox"/> Other	2,281	<input type="checkbox"/> Leibniz Institute for Farm Animal Biology	12
<input type="checkbox"/> 2017	16,285	<input type="checkbox"/> Audiovisual	1,575	<input type="checkbox"/> National Research Institute for Agriculture, Food and Environment	9
<input type="checkbox"/> 2016	12,299	<input type="checkbox"/> Image	1,047	<input type="checkbox"/> University of Clermont Auvergne	8
<input type="checkbox"/> 2015	11,786	<input type="checkbox"/> Software	436	<input type="checkbox"/> Aarhus University	8
<input type="checkbox"/> 2014	265,044	<input type="checkbox"/> Workflow	295	<input type="checkbox"/> Institute for Advanced Sustainability Studies	7
<input type="checkbox"/> 2013	123,527	<input type="checkbox"/> Interactive Resource	112	<input type="checkbox"/> University of Bamberg	5
<input type="checkbox"/> 2012	6,308	<input type="checkbox"/> Sound	30	<input type="checkbox"/> University of Reading	5
		<input type="checkbox"/> Physical Object	27		
		<input type="checkbox"/> Event	24		
		<input type="checkbox"/> Data Paper	18		
		<input type="checkbox"/> Model	6		
		<input type="checkbox"/> Service	3		

Findability

OLAC (Language)



OLAC Language Resource Catalog

Search for language resources

Results:

Showing hits **1 - 50** out of **12598**

Modern theological German : a reader and dictionary -- Theological German
Ziefle, Helmut W., 1939-. 1997. Grand Rapids, Mich. : Baker Books.

The new Cassell's German dictionary : German-English, English-German
Betteridge, Harold T. 1971. New York : Funk & Wagnalls.

The new Cassell's German dictionary : German-English, English German
Betteridge, Harold T. 1962. New York : Funk & Wagnalls Co.

recordings
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

The Berlitz self-teacher: German
Berlitz Schools of Languages of America. 1950. New York : Grosset & Dunlap.

final data
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

recordings
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

AphasiaBank German CAP Corpus
Bates, Elizabeth. 2004-03-30. TalkBank.

German Rigol Corpus
Rigol, Rosemarie. 2007-03-05. TalkBank.

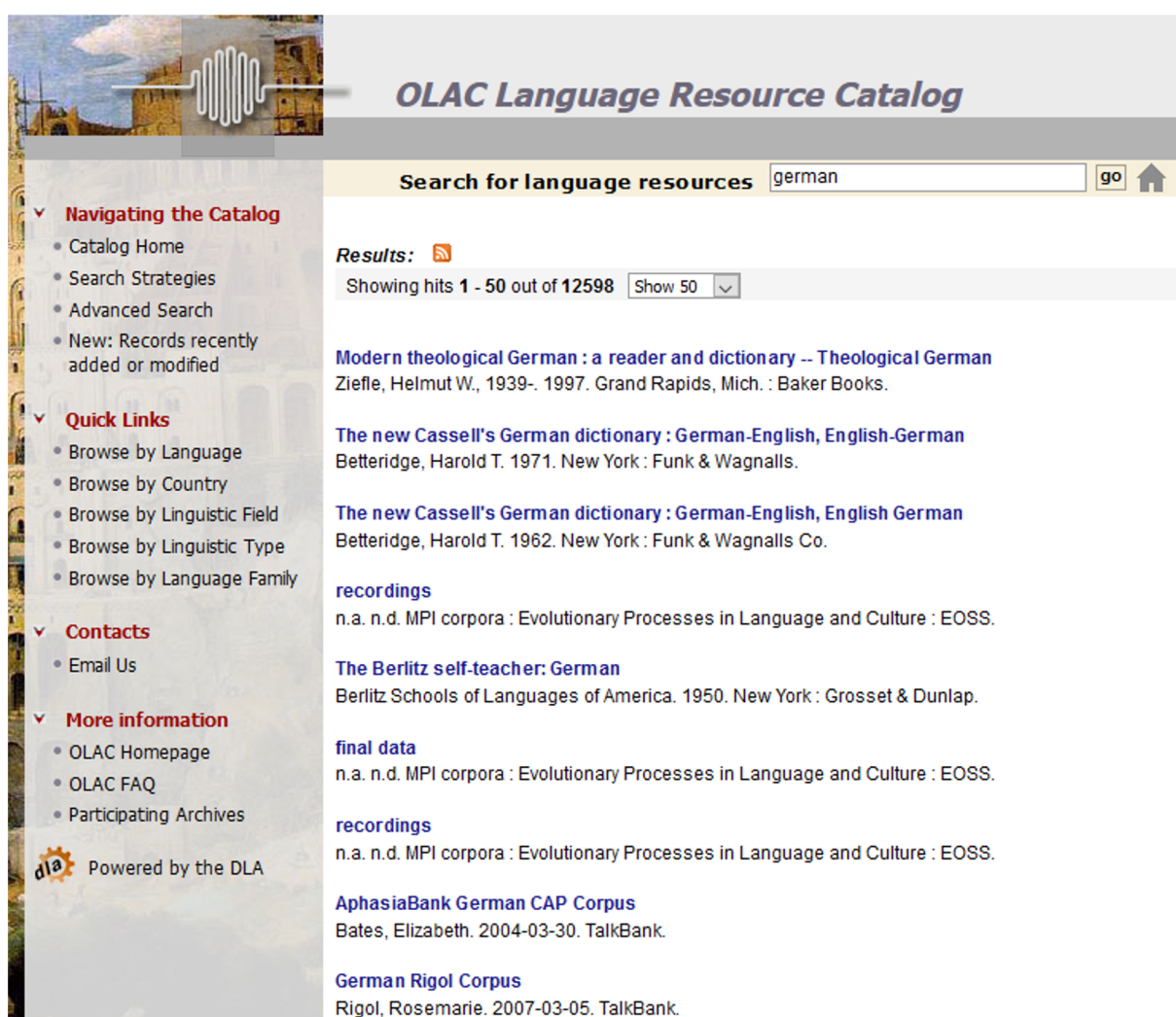
Navigation:

- ▼ **Navigating the Catalog**
 - Catalog Home
 - Search Strategies
 - Advanced Search
 - New: Records recently added or modified
- ▼ **Quick Links**
 - Browse by Language
 - Browse by Country
 - Browse by Linguistic Field
 - Browse by Linguistic Type
 - Browse by Language Family
- ▼ **Contacts**
 - Email Us
- ▼ **More information**
 - OLAC Homepage
 - OLAC FAQ
 - Participating Archives

Powered by the DLA

Findability

OLAC (Language)



OLAC Language Resource Catalog

Search for language resources

Results:

Showing hits **1 - 50** out of **12598**

Modern theological German : a reader and dictionary -- Theological German
Ziefle, Helmut W., 1939-. 1997. Grand Rapids, Mich. : Baker Books.

The new Cassell's German dictionary : German-English, English-German
Betteridge, Harold T. 1971. New York : Funk & Wagnalls.

The new Cassell's German dictionary : German-English, English German
Betteridge, Harold T. 1962. New York : Funk & Wagnalls Co.

recordings
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

The Berlitz self-teacher: German
Berlitz Schools of Languages of America. 1950. New York : Grosset & Dunlap.

final data
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

recordings
n.a. n.d. MPI corpora : Evolutionary Processes in Language and Culture : EOSS.

AphasiaBank German CAP Corpus
Bates, Elizabeth. 2004-03-30. TalkBank.

German Rigol Corpus
Rigol, Rosemarie. 2007-03-05. TalkBank.

Navigation:

- ▼ **Navigating the Catalog**
 - Catalog Home
 - Search Strategies
 - Advanced Search
 - New: Records recently added or modified
- ▼ **Quick Links**
 - Browse by Language
 - Browse by Country
 - Browse by Linguistic Field
 - Browse by Linguistic Type
 - Browse by Language Family
- ▼ **Contacts**
 - Email Us
- ▼ **More information**
 - OLAC Homepage
 - OLAC FAQ
 - Participating Archives

Powered by the DLA

Findability – OLAC (Language)

▼ Subject language	browse
• German	8825
• Swiss German	880
• Polish	625
• German Sign Language	542
view more...	

▼ Content language	browse
• German	9452
• Swiss German	873
• English	863
• Polish	690
view more...	

Findability – OLAC (Language)

ggnheli -- picture story by an adult native speaker of German (control group)

Helga; Thomas Schmalzgrüber; Christine Dimroth (compiler); Thomas Schmalzgrüber (annotator). n.d. MPI corpora : Acquisition : L2 Acquisition : APN.

Vvedenie v altaiskoe ĭazykoznanie

Ramstedt, G. J. (Gustaf John), 1873-1950. n.d. Moskva : Izd-vo inostrannoĭ lit-ry.

imitation-33 -- child learner of L2 German

Christine Dimroth (compiler); Sarah Schimke (compiler). n.d. MPI corpora : Acquisition : L2 Acquisition : Imitation.

imitation-35 -- child learner of L2 German

Christine Dimroth (compiler); Sarah Schimke (compiler). n.d. MPI corpora : Acquisition : L2 Acquisition : Imitation.

Findability – OLAC (Language)

Title: Wvedenie v altaiskoe ĭazykoznanie
Online: No
Archive: [Graduate Institute of Applied Linguistics Library](#) (see archive description)
Contributor: Ramstedt, G. J. (Gustaf John), 1873-1950 (author)
Publisher: Moskva : Izd-vo inostranoĭ lit-ry
Description: Added t.p. in German
Translation of Einführung in die Altaische Sprachwissenschaft

Content language: Russian
German
Subject language: Russian
German

Language family: Indo-European
Germanic
West Germanic

DCMI type: Text

LCSH subject: Altaic languages
Russian language--Translations from German
German language--Translations into Russian

Complete OLAC record: <http://www.language-archives.org/item/oai:gial.edu:16609>

Link for this page: http://dla.library.upenn.edu/dla/olac/record.html?id=gial_edu_16609

Solutions

Solutions

- Improve formats → archives
 - crisp definitions: ~~primary language~~ → subject language, ...
vagueness
 - use non-literals: GlottoCodes, ISO 639-3 codes, ...
ambiguity
- Agree on controlled vocabularies → research community
 - researcher-controlled vocabularies (Glottolog, ORCID, Wikidata)
ambiguity
 - support for vocabularies in tools
- Improve practices → archives, researchers
 - avoid optional fields for core categories in metadata scheme
uncertainty
 - awareness of/following best practices
- Take pivot formats seriously: OLAC → archives
 - only map fitting categories

Thank you

Slides:

<https://tinyurl.com/PARADISECvagueness>

Schwartz, Gabriele, Rau, Felix, and Majka, Nicole 2021. "Vagueness in Metadata" <http://doi.org/10.5281/zenodo.4506936>

Contact (Twitter):

[@fxru](https://twitter.com/fxru) (Felix Rau)

References

- Alexopoulos, P. (2020). Semantic Modeling for Data. O'Reilly: Sebastopol.
- DataCite Metadata Working Group. (2019). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3. DataCite e.V.
- Dickgießer, S. (2009). Metadaten-Schemata in der Datenbank für Gesprochenes Deutsch (DGD 2.0). Institut für Deutsche Sprache Mannheim.
- ISLE Metadata Initiative. (2003). Metadata Elements for Session Descriptions. Version 3.0.4.
- Rau, F. (2018). Basic Language Archive Metadata.
<https://github.com/fxru/blam-metadata>.
- schema.org. <https://schema.org/>

References

- Simons, G., Bird, S., Spanne, J. (2008). OLAC Metadata Usage Guidelines.
- Sorensen, R. (2018). "Vagueness." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2018. Metaphysics Research Lab, Stanford University.

<https://plato.stanford.edu/archives/sum2018/entries/vagueness/>.