A software-driven workflow for the reuse of language documentation data in linguistic studies

Stephan Druskat Kilu von Prince

Dept. of German Studies and Linguistics, Humboldt-Universität zu Berlin, Germany

3rd Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, 27 February 2019

Outline

1. Identify commonplace issues with reuse of language documentation data
2. Describe a workflow based on the use of a specific software toolkit
3. Describe implementation use case

Challenges in the reuse of language documentation data

Reuse of language documentation data

- Typological/cross-linguistic studies rely on corpus data
- Different sources (archives, fieldwork, data exchange)

Two main challenges

- Evaluation of suitability for research question
- Data processing (evaluation, annotation, analysis)
  - Different data formats

Suitability evaluation of a data set

- Includes fine-grained assessment (e.g., occurrence of linguistic phenomena)
- Metadata as potential source of information
  - FAIR metadata?
    "meta(data) [that are] richly described with a plurality of accurate and relevant attributes", or "(meta)data [that do not] meet domain-relevant community standards" (Wilkinson et al., 2016, p.4)
  - Search capabilities in archive interfaces?
- Data exploration
  - can be expensive

Handling different data formats

Evaluation of data sets

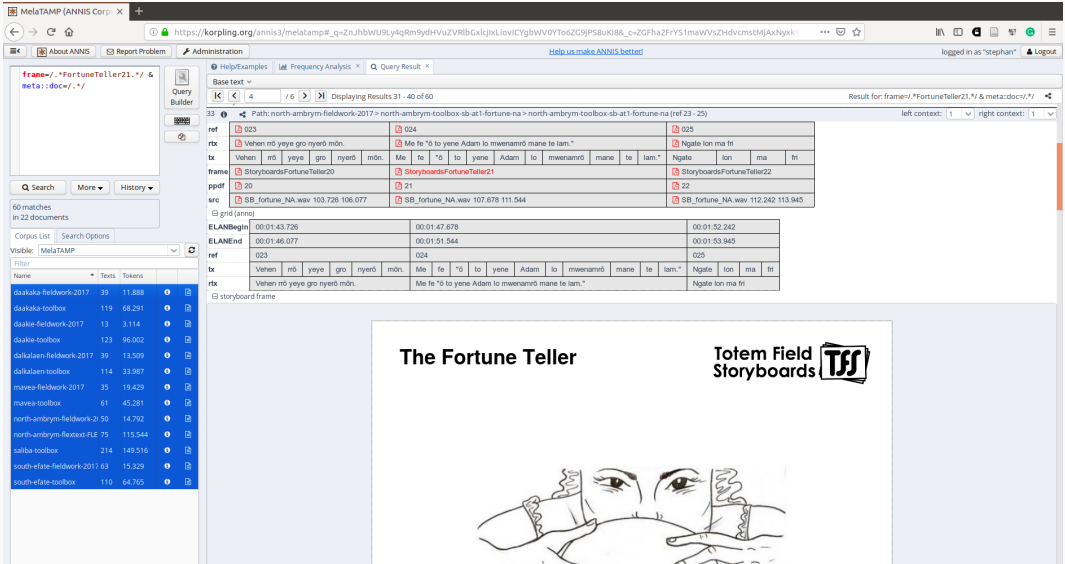
- Different data formats = different tools?
- Tools suitable for analysis? (Can I search at all?)
- Analysis comparability across corpora from different formats? (Can I compare the results of the search?)
- Analysis compatibility: can same features be queried across formats/corpora? (Can I search for the same thing?)

A software-driven workflow

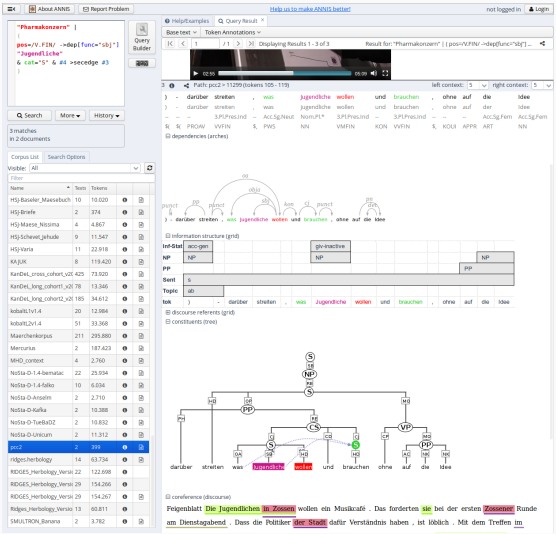
Strategy for solving evaluation and format issues

- Use a software for fine-grained data exploration
  (ANNIS, Krause and Zeldes, 2016)
  - Bypass missing/incomplete metadata
  - Reduce cost of data exploration
- Automatically convert between formats
  (Pepper, Zipser et al., 2011)
- corpus-tools.org
  (Apache License, Version 2.0)

ANNIS GUI (language documentation data)



ANNIS GUI (multi-layer corpus data)



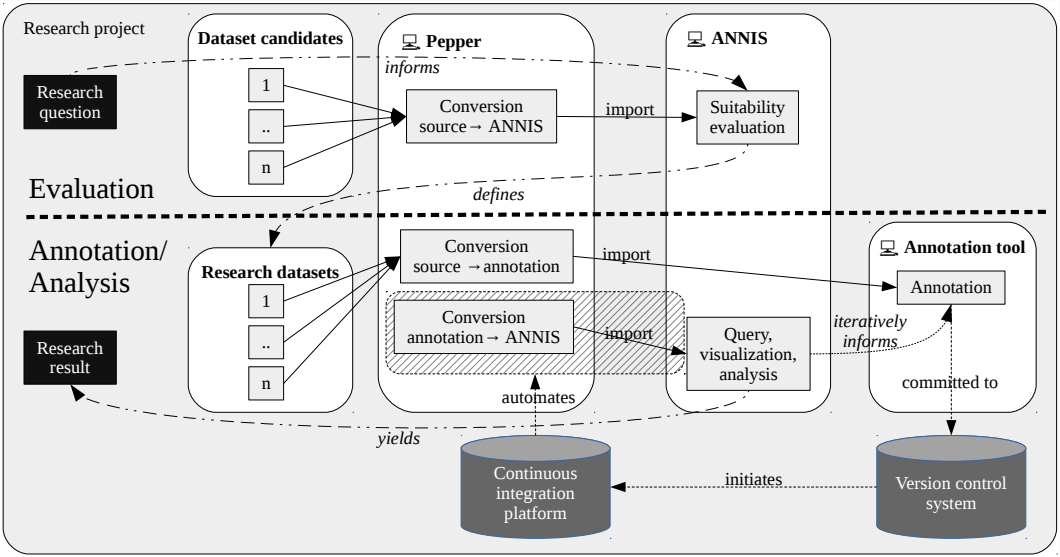
Requirements for evaluation in ANNIS

- Evaluation based on knowledge about existing annotations
  - ANNIS provides an overview of existing annotations and example queries
- Corpora must exist in the ANNIS format
  - (Automatable) conversion solves the data formats issue

Workflow

- (Typological) research may need additional annotation
- Workflow includes:
  1. Evaluation of candidate data sets for study or single research questions
  2. Compilation of study data sets
  3. Annotation
  4. Analysis
  5. Formulation of research results
- Iterations of 3.–5.

Workflow


Research project
Dataset candidates
Pepper
ANNIS
Research question
informs
1
..
n
Conversion source → ANNIS
import
Suitability evaluation
Evaluation
defines
Annotation/ Analysis
Research datasets
Conversion source → annotation
import
Annotation tool
Annotation
1
..
n
Conversion annotation → ANNIS
import
Query, visualization, analysis
iteratively informs
Research result
automates
committed to
yields
Continuous integration platform
initiates
Version control system

Case study: Workflow implementation in the MelaTAMP project

MelaTAMP

- 2016–2019, DFG-funded (grant no. 273640553)
- M. Krifka (PI), K. von Prince (PI), A. Krajinovic (Researcher), me (RSE), A. Tjuka & L. Weißmann (student assistants)
- Research question: *TAMP systems in Melanesian (mood-prominent) languages*
- Data sets: 13 corpora in 3 formats (Toolbox, ELAN, FLEx) from 7 researchers (project staff and collaborators)
- RDM: HZSK repository & PARADISEC
  - HZSK prefers deposit in EXMARaLDA format
  - ANNIS also available

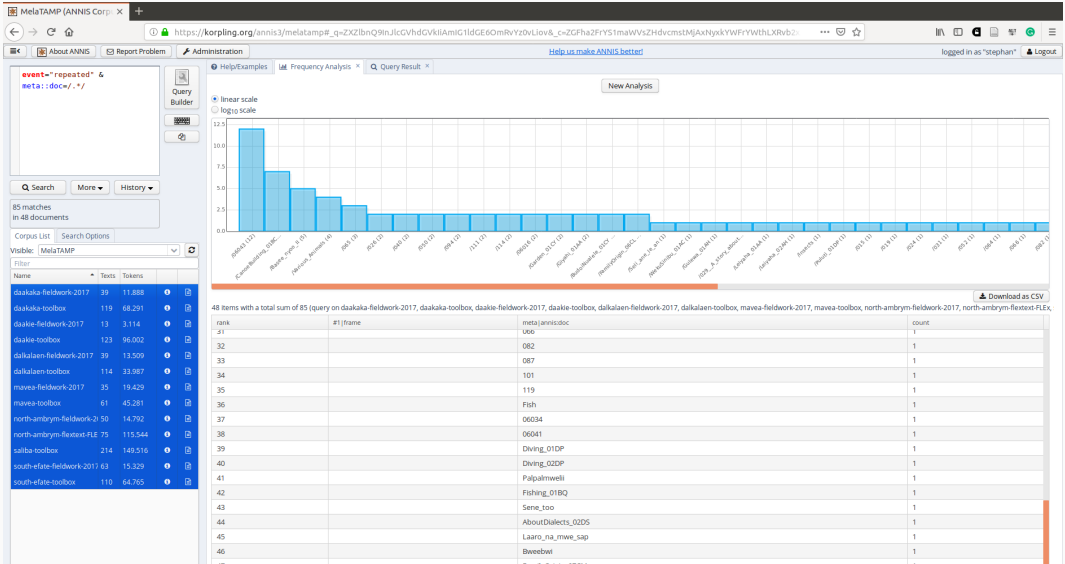
Corpora

Language	ISO 639-3	Tokens	Country	Elicitator	Format (Software)
Daakie	ptv	~86k	Vanuatu	Krifka (2013)	Text (Toolbox)
Daakie	ptv	~3k	Vanuatu	Manfred Krifka	Text (Toolbox)
Daakaka	bpa	~59k	Vanuatu	von Prince (2013a)	Text (Toolbox)
Daakaka	bpa	~80k	Vanuatu	Kilu von Prince	XML (ELAN)
Dalkalaen		~30k	Vanuatu	von Prince (2013b)	Text (Toolbox)
Dalkalaen		~13k	Vanuatu	Kilu von Prince	XML (ELAN)
North Ambrym	mmg	~24k	Vanuatu	Franjieh (2013)	XML (FLEx)
North Ambrym	mmg	~15k	Vanuatu	Michael Franjieh	XML (ELAN)
Mavea	mkv	~30k	Vanuatu	Guérin (2006)	Text (Toolbox)
Mavea	mkv	~12k	Vanuatu	Valérie Guérin	Text (Toolbox)
South Efate	erk	~54k	Vanuatu	Thieberger (2006)	Text (Toolbox)
South Efate	erk	~15k	Vanuatu	Ana Krajinovic	XML (ELAN)
Saliba/Logea	sbe	~138k	PNG	Margetts et al. (2017)	Text (Toolbox)

Evaluation

- Controlled to some extent, no unbound evaluation (own data, newly elicited, direct exchange)
- Evaluation for different detailed research questions in ANNIS
  - E.g., “For a study of habitual contexts of repetition events, which corpora contain repetition events?”

Evaluation


MelaTAMP (ANNIS Corp
https://korpling.org/annis3/melatamp#_q=ZXZlbnQ9InJlcGVhdGVkIiAmIG1ldGE6OmRvYzo0Liov&_c=ZGFhaZFrYS1maWVsZHdvcmstMjAxNyxkYWFrYWthLXRvb2
About ANNIS
Report Problem
Administration
Help us make ANNIS better!
logged in as "stephan"
Logout
event="repeated" &
meta::doc=/.*/
Query Builder
Search
More
History
85 matches
in 48 documents
Corpus List
Search Options
Visible:
MelaTAMP
Filter
Name
Texts
Tokens
daakaka-fieldwork-2017 39 11.888
daakaka-toolbox 119 68.291
daakie-fieldwork-2017 13 3.114
daakie-toolbox 123 96.002
dalkalaen-fieldwork-2017 39 13.509
dalkalaen-toolbox 114 33.987
mavea-fieldwork-2017 35 19.429
mavea-toolbox 61 45.281
north-ambrym-fieldwork-2(50 14.792
north-ambrym-flextext-FLE 75 115.544
saliba-toolbox 214 149.516
south-efate-fieldwork-2017 63 15.329
south-efate-toolbox 110 64.765
Help/Examples
Frequency Analysis
Query Result
New Analysis
linear scale
log10 scale
12.5
10.0
7.5
5.0
2.5
0.0
Download as CSV
48 items with a total sum of 85 (query on daakaka-fieldwork-2017, daakaka-toolbox, daakie-fieldwork-2017, daakie-toolbox, dalkalaen-fieldwork-2017, dalkalaen-toolbox, mavea-fieldwork-2017, mavea-toolbox, north-ambrym-fieldwork-2017, north-ambrym-flextext-FLEx,
rank
#1 | frame
meta | annis:doc
count
32 082 1
33 087 1
34 101 1
35 119 1
36 Fish 1
37 06034 1
38 06041 1
39 Diving_01DP 1
40 Diving_02DP 1
41 Palpalmweli 1
42 Fishing_01BQ 1
43 Sene_too 1
44 AboutDialects_02DS 1
45 Laaro_na_mwe_sap 1
46 Bweebwi 1

Annotation

- Toolbox interlinear text
  - (+) Researcher preference
  - (+) RegEx
  - (+) Human-readability & quick manual annotation
  - (–) Underdefined
  - (–) Only most basic validation possible

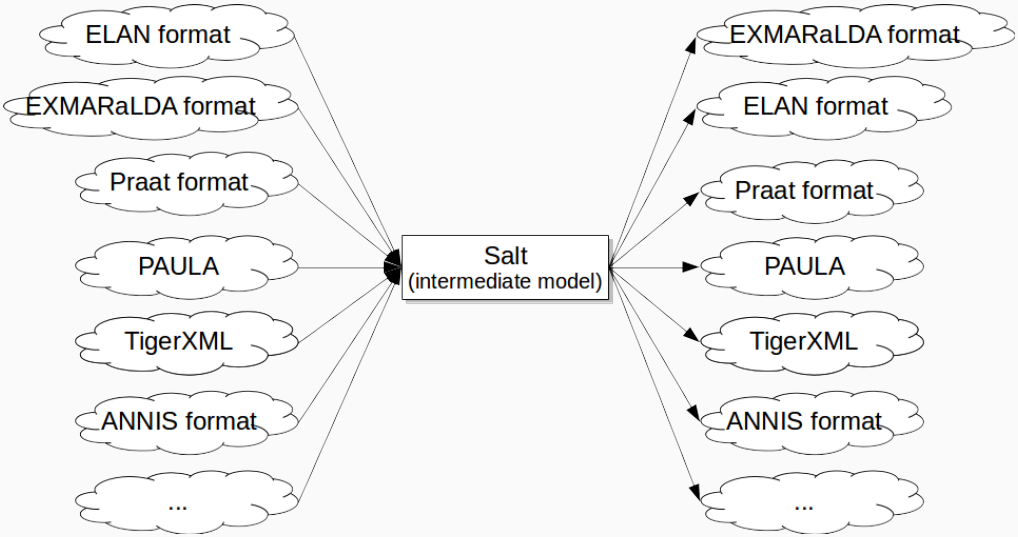
```
\_sh v3.0 400 Text
\id {Document}
\some_marker {e.g., metadata}

\ref {phrase}
\tx Lexical information
\mb Morphological information
```

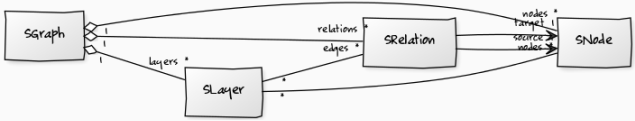
Conversion process

- Pepper: CLI tool taking workflow decriptions as input
- Enables n:n conversion via intermediate model, a Salt instance (Zipser and Romary, 2010)
- Salt is a versatile graph-based meta-model for linguistic data

Pepper


ELAN format
EXMARaLDA format
Praat format
PAULA
TigerXML
ANNIS format
...
Salt
(intermediate model)
EXMARaLDA format
ELAN format
Praat format
PAULA
TigerXML
ANNIS format
...

Salt


SGraph
SRelation
SNode
SLayer
nodes *
target 1
source 1
nodes 2
relations *
edges *
layers *
1
1
1
*
*

Example: Toolbox import (ToolboxTextModules)

Druskat (2018b)

- Tokenization
- Span building
- Normalization
- Index-based sub-spanning (not supported in Toolbox)
- Detection & “fixing” of interlinearization errors

Conversion configuration

```
<?xml version='1.0' encoding='UTF-8'?>
<pepper-job id="daakaka-toolbox-corpus-toolbox-to-annis" version="1.0">
    <importer name="ToolboxTextImporter" path="../toolbox-corpus/toolbox/">
        <property key="subrefDefinitionMarker">subref</property>
        <property key="subrefAnnotationMarkers">clause, time, mood, event, polarity</property>
        <property key="normalizeMarkers">true</property>
    </importer>
    <manipulator name="OrderRelationAdder">
        <customization>
            <property key="segmentation-layers">{ref}</property>
        </customization>
    </manipulator>
    <exporter name="ANNISExporter" path="../toolbox-corpus/annis/">
        <property key="clobber.visualisation">false</property>
        <property key="corpusName">daakaka-toolbox</property>
    </exporter>
</pepper-job>
```

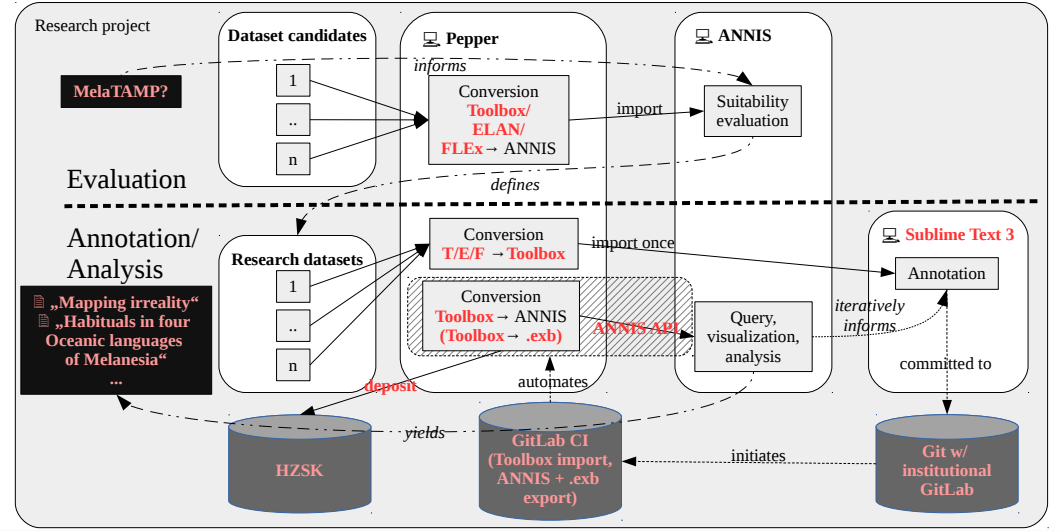
Workflow

1. Convert corpora to Toolbox format
   - Manual ELAN export
   - Pepper FLExModules (Druskat, 2018a) import + Pepper ToolboxTextModules (Druskat, 2018b) export
2. Convert from Toolbox format to ANNIS & upload to ANNIS
   - Pepper ToolboxTextModules import + Pepper ANNISModules (pepperModules-ANNISModules Team, 2018) export
3. Convert from Toolbox to EXMARaLDA `.exb` and save for deposit
   - Pepper ToolboxTextModules import + Pepper EXMARaLDAModules (pepperModules-EXMARaLDAModules Team, 2018) export

Reproducible conversion

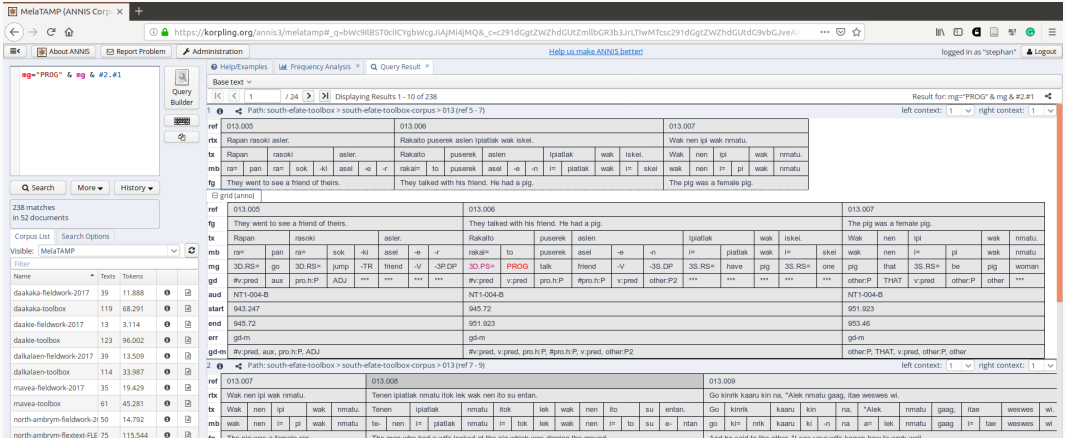
- Conversion paths in Pepper can be persisted in XML for reproducible conversion
- Speeds up the conversion process, enables automation
- Enables conversion testing (implicitly implemented in module unit tests)
- Enables “continuous analysis” (Beaulieu-Jones and Greene, 2017) via automated CI whenever annotations change

Concrete workflow



Analysis example

(?) “Which subject markers can the progressive marker in Nafsan combine with?”



Analysis example

(!) “The progressive marker in Nafsan (PROG) can combine with realis subject markers (RS), perfect subject markers (PS) and irrealis subject markers (IRS), i.e., some markers have restrictions on subject markers they combine with. The analysis also shows some other combinations like ‘still’, ‘unable’ and ‘every’ which come in between the subject marking and the verb.”

Conclusion

Conclusion

- ANNIS for data exploration and corpus queries
- Implementation of the (continuous) annotation–analysis part of the proposed workflow in MelaTAMP
- Enablement of evaluation through the provision of new Pepper modules
  - Potential for circumvention of obstacles to data reuse of non-FAIR data sets
- Automation enabled us to efficiently analyse expressions of irrealis and habitual contexts across our corpora
  (von Prince et al., 2019, von Prince et al., forthcoming)

Thank you!

Mahalo!

Questions?

- hu.berlin/melatamp
- corpus-tools.org
- sdruskat.net
- stephan.druskat@hu-berlin.de
- Twitter: @stdruskat

References

- Anna Margetts, Andrew Margetts, and Carmen Dawuda. 2017. Saliba/Logea. The Language Archive. http://dobes.mpi.nl/projects/saliba.
- Brett K. Beaulieu-Jones and Casey S. Greene. 2017. Reproducibility of computational workflows is automated using continuous analysis. Nature biotechnology, 35(4):342–346, April. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
- Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In Proceedings of the Workshop on Language Resource and Language Technology Standards.
- Florian Zipser, Amir Zeldes, Julia Ritz, Laurent Romary, and Ulf Leser. 2011. Pepper: Handling a multiverse of formats. https://doi.org/10.5281/zenodo.15638
- Kilu von Prince, Ana Krajinović, Anna Margetts, Nick Thieberger, and Valérie Guérin. 2019. Habituality in four Oceanic languages of Melanesia. STUF - Language Typology and Universals, 72(1):21–66. https://doi.org/10.1515/stuf-2019-0002
- Kilu von Prince, Ana Krajinović, Manfred Krifka, Valérie Guérin, and Michael Franjieh. forthcoming. Mapping irreality: Storyboards for eliciting TAM contexts. In Proceedings of Linguistic Evidence 2018.
- Kilu von Prince. 2013a. Daakaka, The Language Archive. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.
- Kilu von Prince. 2013b. Dalkalaen, The Language Archive. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.
- Manfred Krifka. 2013. Daakie, The Language Archive. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, March.

References cont.

- Michael Franjieh. 2013. A documentation of North Ambrym, a language of Vanuatu. SOAS, Endangered Languages Archive. https://elar.soas.ac.uk/Collection/MPI67426. [Accessed on 2017/10/04], London.
- Nick Thieberger. 2006. Dictionary and texts in South Efate. Digital collection managed by PARADISEC. DOI: https://doi.org/10.4225/72/56FA0C5A7C98F.
- pepperModules-ANNISModules Team. 2018. pepperModules-ANNISModules (Version 2.0.9). Software Heritage. https://archive.softwareheritage.org/swh:1:rev:2d51c9045259fd92734db8cec0732928232c5d12;origin=https://github.com/korpling/pepperModules-ANNISModules/
- pepperModules-EXMARaLDAModules Team. 2018. pepperModules-EXMARaLDAModules (Version 1.2.2). GitHub. https://github.com/korpling/pepperModules-EXMARaLDAModules/releases/tag/pepperModules-EXMARaLDAModules-1.2.2
- Stephan Druskat. 2018a. FLExModules (Version 1.0.8). Zenodo, December. https://doi.org/10.5281/zenodo.2247370
- Stephan Druskat. 2018b. ToolboxTextModules (Version 1.0.0). Zenodo, January. https://doi.org/10.5281/zenodo.1162207
- Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities, 31(1):118–139. https://doi.org/10.1093/llc/fqu057
- Valérie Guérin. 2006. Mavea. SOAS, Documentation of Endangered Languages Archive. https://elar.soas.ac.uk/Collection/MPI67426. [Accessed on 2017/03/01], London.