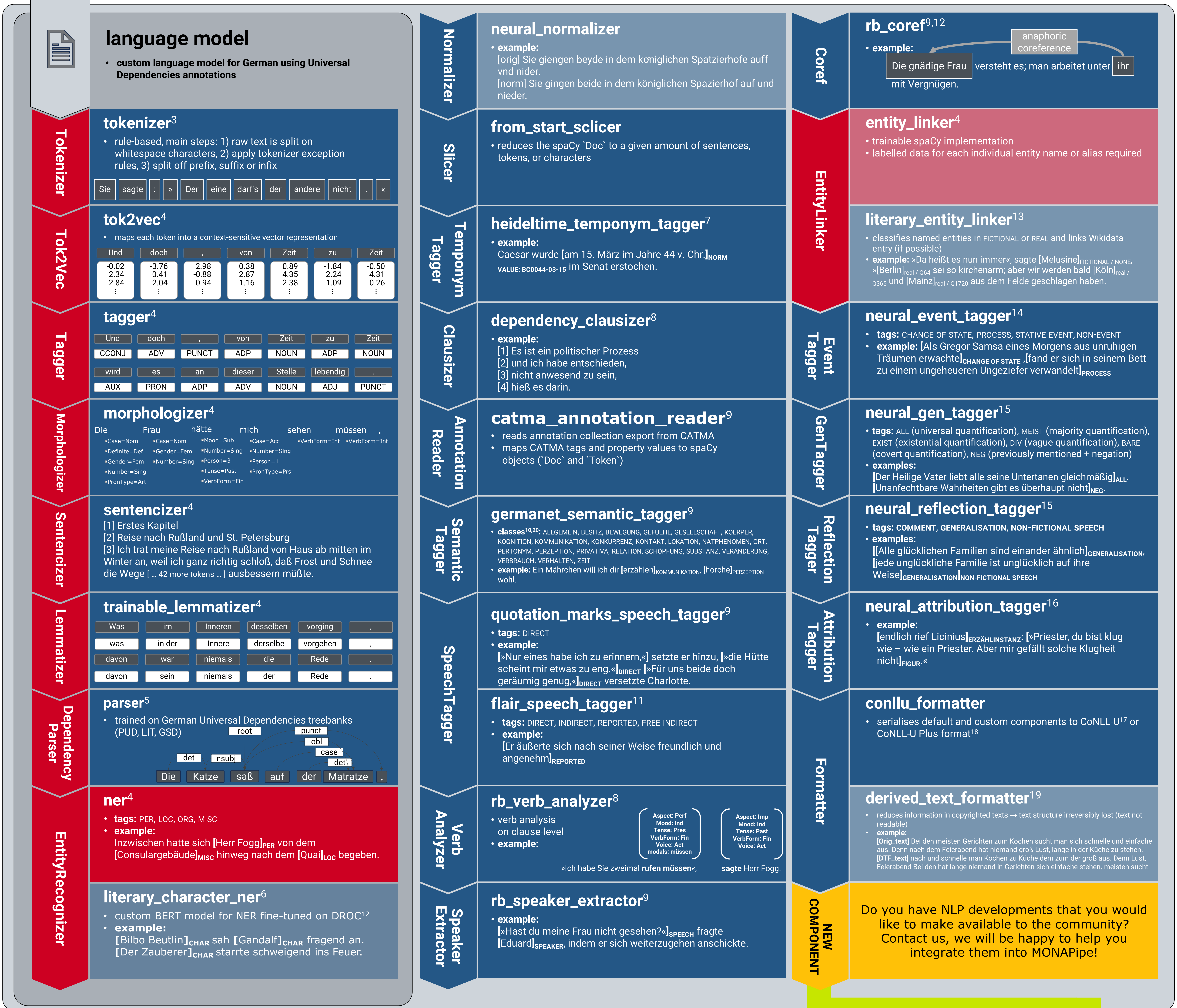
MONAPipe: Modular Natural Language Processing Pipeline for Digital Humanities

Text+

input:
text file

- natural language processing pipeline based on Python library spaCy[1]
- jointly developed by partners in Text+ task area Collections as a software service
- originally created in the project „Modes of Narration and Attribution“ (MONA)[2]
- part of the Text+ data processing pipelines portfolio
- with involvement of the Text+ task area Infrastructure Operations
- upcoming: Python Package and integration into Text+ infrastructure

custom component	implementation 1
	implementation 2 (transparency indicates implementation is in development or requires training)
default component	implementation



language model

- custom language model for German using Universal Dependencies annotations

Tokenizer

tokenizer[3]

- rule-based, main steps: 1) raw text is split on whitespace characters, 2) apply tokenizer exception rules, 3) split off prefix, suffix or infix

Sie	sagte	:	»	Der	eine	darf's	der	andere	nicht	.	«

Tok2Vec

tok2vec[4]

- maps each token into a context-sensitive vector representation

Und	doch	,	von	Zeit	zu	Zeit
-0.02 2.34 2.84 ⋮	-3.76 0.41 2.04 ⋮	2.98 -0.88 -0.94 ⋮	0.38 2.87 1.16 ⋮	0.89 4.35 2.38 ⋮	-1.84 2.24 -1.09 ⋮	-0.50 4.31 -0.26 ⋮

Tagger

tagger[4]

Und	doch	,	von	Zeit	zu	Zeit
CCONJ	ADV	PUNCT	ADP	NOUN	ADP	NOUN
wird	es	an	dieser	Stelle	lebendig	.
AUX	PRON	ADP	ADV	NOUN	ADJ	PUNCT

Morphologizer

morphologizer[4]

Die	Frau	hätte	mich	sehen	müssen	.
Case=Nom Definite=Def Gender=Fem Number=Sing PronType=Art	Case=Nom Gender=Fem Number=Sing	Mood=Sub Number=Sing Person=3 Tense=Past VerbForm=Fin	Case=Acc Number=Sing Person=1 PronType=Prs	VerbForm=Inf	VerbForm=Inf	

Sentencizer

sentencizer[4]

[1] Erstes Kapitel
[2] Reise nach Rußland und St. Petersburg
[3] Ich trat meine Reise nach Rußland von Haus ab mitten im Winter an, weil ich ganz richtig schloß, daß Frost und Schnee die Wege [… 42 more tokens …] ausbessern müßte.

Lemmatizer

trainable_lemmatizer[4]

Was	im	Inneren	desselben	vorging	,
was	in der	Innere	derselbe	vorgehen	,
davon	war	niemals	die	Rede	.
davon	sein	niemals	der	Rede	.

Dependency Parser

parser[5]

- trained on German Universal Dependencies treebanks (PUD, LIT, GSD)

EntityRecognizer

ner[4]

- **tags:** PER, LOC, ORG, MISC
- **example:**
  Inzwischen hatte sich [Herr Fogg]$_{\text{PER}}$ von dem [Consulargebäude]$_{\text{MISC}}$ hinweg nach dem [Quai]$_{\text{LOC}}$ begeben.

literary_character_ner[6]

- custom BERT model for NER fine-tuned on DROC[12]
- **example:**
  [Bilbo Beutlin]$_{\text{CHAR}}$ sah [Gandalf]$_{\text{CHAR}}$ fragend an. [Der Zauberer]$_{\text{CHAR}}$ starrte schweigend ins Feuer.

Normalizer

neural_normalizer

- **example:**
  [orig] Sie giengen beyde in dem koniglichen Spatzierhofe auff vnd nider.
  [norm] Sie gingen beide in dem königlichen Spazierhof auf und nieder.

Slicer

from_start_sclicer

- reduces the spaCy `Doc` to a given amount of sentences, tokens, or characters

Temponym Tagger

heideltime_temponym_tagger[7]

- **example:**
  Caesar wurde [am 15. März im Jahre 44 v. Chr.]$_{\text{NORM}}$ VALUE: BC0044-03-15 im Senat erstochen.

Clausizer

dependency_clausizer[8]

- **example:**
  [1] Es ist ein politischer Prozess
  [2] und ich habe entschieden,
  [3] nicht anwesend zu sein,
  [4] hieß es darin.

Annotation Reader

catma_annotation_reader[9]

- reads annotation collection export from CATMA
- maps CATMA tags and property values to spaCy objects (`Doc` and `Token`)

Semantic Tagger

germanet_semantic_tagger[9]

- **classes[10,20]:** ALLGEMEIN, BESITZ, BEWEGUNG, GEFUEHL, GESELLSCHAFT, KOERPER, KOGNITION, KOMMUNIKATION, KONKURRENZ, KONTAKT, LOKATION, NATPHENOMEN, ORT, PERTONYM, PERZEPTION, PRIVATIVA, RELATION, SCHÖPFUNG, SUBSTANZ, VERÄNDERUNG, VERBRAUCH, VERHALTEN, ZEIT
- **example:** Ein Mährchen will ich dir [erzählen]$_{\text{KOMMUNIKATION}}$, [horche]$_{\text{PERZEPTION}}$ wohl.

SpeechTagger

quotation_marks_speech_tagger[9]

- **tags:** DIRECT
- **example:**
  [»Nur eines habe ich zu erinnern,«] setzte er hinzu, [»die Hütte scheint mir etwas zu eng.«]$_{\text{DIRECT}}$ [»Für uns beide doch geräumig genug,«]$_{\text{DIRECT}}$ versetzte Charlotte.

flair_speech_tagger[11]

- **tags:** DIRECT, INDIRECT, REPORTED, FREE INDIRECT
- **example:**
  [Er äußerte sich nach seiner Weise freundlich und angenehm]$_{\text{REPORTED}}$

Verb Analyzer

rb_verb_analyzer[8]

- verb analysis on clause-level
- **example:**

Aspect: Perf
Mood: Ind
Tense: Pres
VerbForm: Fin
Voice: Act
modals: müssen

Aspect: Imp
Mood: Ind
Tense: Past
VerbForm: Fin
Voice: Act

»Ich habe Sie zweimal **rufen müssen**«, **sagte** Herr Fogg.

Speaker Extractor

rb_speaker_extractor[9]

- **example:**
  [»Hast du meine Frau nicht gesehen?«]$_{\text{SPEECH}}$ fragte [Eduard]$_{\text{SPEAKER}}$, indem er sich weiterzugehen anschickte.

Coref

rb_coref[9,12]

- **example:**

Die gnädige Frau versteht es; man arbeitet unter ihr mit Vergnügen.

EntityLinker

entity_linker[4]

- trainable spaCy implementation
- labelled data for each individual entity name or alias required

literary_entity_linker[13]

- classifies named entities in FICTIONAL or REAL and links Wikidata entry (if possible)
- **example:** »Da heißt es nun immer«, sagte [Melusine]$_{\text{FICTIONAL / NONE}}$, »[Berlin]$_{\text{real / Q64}}$ sei so kirchenarm; aber wir werden bald [Köln]$_{\text{real / Q365}}$ und [Mainz]$_{\text{real / Q1720}}$ aus dem Felde geschlagen haben.

Event Tagger

neural_event_tagger[14]

- **tags:** CHANGE OF STATE, PROCESS, STATIVE EVENT, NON-EVENT
- **example:** [Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte]$_{\text{CHANGE OF STATE}}$,[fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt]$_{\text{PROCESS}}$

GenTagger

neural_gen_tagger[15]

- **tags:** ALL (universal quantification), MEIST (majority quantification), EXIST (existential quantification), DIV (vague quantification), BARE (covert quantification), NEG (previously mentioned + negation)
- **examples:**
  [Der Heilige Vater liebt alle seine Untertanen gleichmäßig]$_{\text{ALL}}$.
  [Unanfechtbare Wahrheiten gibt es überhaupt nicht]$_{\text{NEG}}$.

Reflection Tagger

neural_reflection_tagger[15]

- **tags:** COMMENT, GENERALISATION, NON-FICTIONAL SPEECH
- **examples:**
  [[Alle glücklichen Familien sind einander ähnlich]$_{\text{GENERALISATION}}$, [jede unglückliche Familie ist unglücklich auf ihre Weise]$_{\text{GENERALISATION}}$]$_{\text{NON-FICTIONAL SPEECH}}$

Attribution Tagger

neural_attribution_tagger[16]

- **example:**
  [endlich rief Licinius]$_{\text{ERZÄHLINSTANZ}}$: [»Priester, du bist klug wie – wie ein Priester. Aber mir gefällt solche Klugheit nicht]$_{\text{FIGUR}}$.«

Formatter

conllu_formatter

- serialises default and custom components to CoNLL-U[17] or CoNLL-U Plus format[18]

derived_text_formatter[19]

- reduces information in copyrighted texts → text structure irreversibly lost (text not readable)
- **example:**
  [Orig_text] Bei den meisten Gerichten zum Kochen sucht man sich schnelle und einfache aus. Denn nach dem Feierabend hat niemand groß Lust, lange in der Küche zu stehen.
  [DTF_text] nach und schnelle man Kochen zu Küche dem zum der groß aus. Denn Lust, Feierabend Bei den hat lange niemand in Gerichten sich einfache stehen. meisten sucht

NEW COMPONENT

Do you have NLP developments that you would like to make available to the community? Contact us, we will be happy to help you integrate them into MONAPipe!

output:
spaCy Doc

References

1) Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303.
2) https://www.uni-goettingen.de/en/mona/626918.html.
3) https://spacy.io/usage/linguistic-features#how-tokenizer-works.
4) For components/implementations from spaCy see: https://spacy.io/api.
5) Tillmann Dönicke. 2023. German UD spaCy model (md), https://doi.org/10.25625/S2LPJP, GRO.data, V1.
6) severinsimmler/literary-german-bert, https://huggingface.co/severinsimmler/literary-german-bert.
7) Jannik Strötgen and Michael Gertz. 2015. A baseline temporal tagger for all languages. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, S. 541–547, Lisbon, Portugal. Association for Computational Linguistics.
8) Tillmann Dönicke. 2020. Clause-level tense, mood, voice and modality tagging for German. In Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories, S. 1–17, Düsseldorf, Germany. Association for Computational Linguistics.
9) Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy. In Proceedings of KONVENS (Konferenz zur Verarbeitung natürlicher Sprache/Conference on Natural Language Processing).
10) Beth Levin. 1995. English verb classes and alternations. A preliminary Investigation, 1.
11) Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer und Fotis Jannidis. 2020. To BERT or not to BERT – comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS).
12) Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, S. 98–104, Denver, Colorado, USA. Association for Computational Linguistics.
13) Florian Barth, Hanna Varachkina, Tillmann Dönicke und Luisa Gödeke. 2022. Levels of non-fictionality in fictional texts. In Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022, S. 27–32, Marseille, France. European Language Resources Association.
14) Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021), S. 333–345, Amsterdam, the Netherlands.
15) Thorben Schomacker, Tillmann Dönicke, and Marina Tropmann-Frick (2022). Automatic Identification of Generalizing Passages in German Fictional Texts using BERT with Monolingual and Multilingual Training Data. Extended abstract submitted and accepted for the KONVENS 2022 Student Poster Session.
16) Tillmann Dönicke, Hanna Varachkina, Anna Mareike Weimer, Luisa Gödeke, Florian Barth, Benjamin Gittel, Anke Holler, and Caroline Sporleder. 2022. Modelling speaker attribution in narrative texts with biased and bias-adjustable neural networks. Frontiers in Artificial Intelligence, 4.
17) https://universaldependencies.org/format.html.
18) https://universaldependencies.org/ext-format.html.
19) Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. Abgeleitete Textformate: Text and Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel. text/html Format. DOI: 10.17175/2020_006.
20) Franz Hundsnurscher and Jochen Splett. 1982. Semantik der Adjektive des Deutschen. Analyse der semantischen Relationen.

Text+ Plenary: Connecting People and Data
28./29. September 2023

https://gitlab.gwdg.de/text-plus-collections/mona-pipe
Contact: florian.barth@uni-goettingen.de

Florian Barth (SUB), Yannic Bracke (BBAW),
José Calvo Tello (SUB), George Dogaru (GWDG),
Tillmann Dönicke (SUB), Keli Du (TCDH),
Stefan E. Funk (SUB), Philippe Genet (DNB),
Mathias Göbel (SUB), Lennart Keller (UniWü),
Daniel Kurzawe (SUB), Ubbo Veentjer (SUB),
Lukas Weimer (SUB)

Kompetenzzentrum Trier Center for Digital Humanities

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

DEUTSCHE NATIONAL BIBLIOTHEK

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN SUB