

# Identifikation und Disambiguierung von Ortsbezeichnungen in bibliografischen Metadaten auf Grundlage von maschinellem Lernen zur Anreicherung mit Koordinaten



- |             |               |            |              |             |                |             |               |
|-------------|---------------|------------|--------------|-------------|----------------|-------------|---------------|
| ● Arlington | ● Burlington  | ● Clinton  | ● Franklin   | ● Jackson   | ● Manchester   | ● Newport   | ● Salem       |
| ● Ashland   | ● Centerville | ● Dayton   | ● Georgetown | ● Kingston  | ● Milford      | ● Oakland   | ● Springfield |
| ● Auburn    | ● Clayton     | ● Dover    | ● Greenville | ● Lexington | ● Milton       | ● Oxford    | ● Washington  |
| ● Bristol   | ● Cleveland   | ● Fairview | ● Hudson     | ● Madison   | ● Mount Vernon | ● Riverside | ● Winchester  |



[1]

○ East London, Südafrika

📍 Berlin, 5660, Südafrika

+ Reiseziel hinzufügen

Jetzt starten ▾

OPTIONEN

📱 Wegbeschreibung an mein Smartphone senden

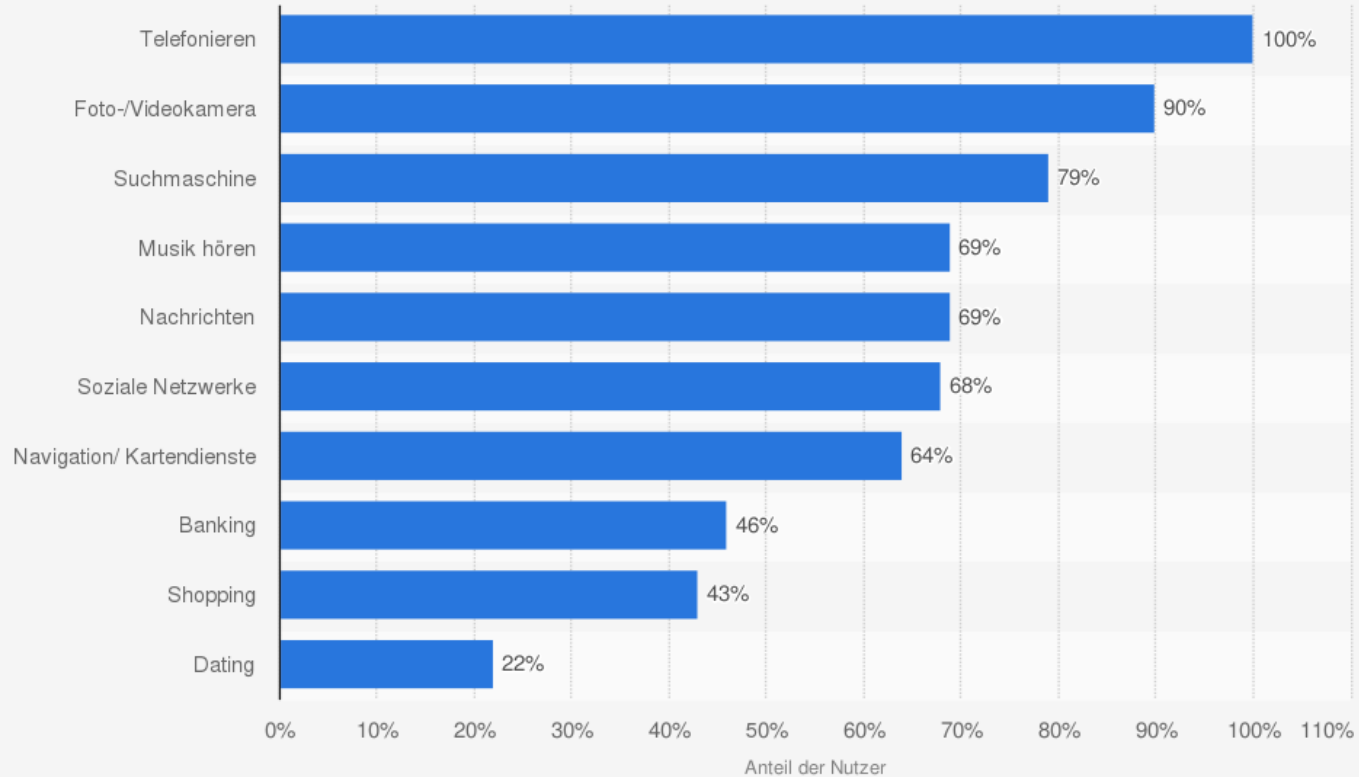
🚗 über N2 **46 min**

Schnellste Route; übliche Verkehrslage 53,2 km

DETAILS



## Anteil der befragten Smartphone-Nutzer, die die folgenden Funktionen mit ihrem Smartphone nutzen



Quelle  
Bitkom  
© Statista 2021

Weitere Informationen:  
Deutschland; Bitkom Research; Januar 2017; Smartphone-Nutzer; ab 14 Jahre

Management Alle Felder Suchen [Erweitert](#)

[Filter zurücksetzen](#) Geografische Suche: **Intersects(ENVELOPE(-134.208984375, -95.712890625, 56.2954074618**

Suchergebnisse

[Suchbereich definieren](#) [Hilfe](#)



Leaflet | Map tiles: Stamen Design | Data: OpenStreetMap



=LDR 03747nz a2200709nc 4500  
=001 040182665  
=003 DE-101  
=005 20181113150134.0  
=008 880701n||azznnaabn\\\\\\\\\\\\\\\\\\|ana\\\\\\\\|c  
=024 7\\\$a4018266-6\$0http://d-nb.info/gnd/4018266-6\$2gnd  
=034 \\\$dE 013 20 19\$eE 013 20 19\$N 050 54 39\$gN 050 54 39\$2geonames\$0http://sws.geonames.org/2925192\$9A:agx  
=034 \\\$dE013.338611\$eE013.338611\$fN050.910833\$gN050.910833\$2geonames\$0http://sws.geonames.org/2925192\$9A:dgx  
=035 \\\$a(DE-101)040182665  
=035 \\\$a(DE-588)4018266-6  
=035 \\\$z(DE-588)2010860-6  
=035 \\\$z(DE-588b)6519955-8  
=035 \\\$z(DE-588b)2010860-6\$9v:zg  
=035 \\\$z(DE-588c)4018266-6\$9v:zg  
=040 \\\$aDE-101\$cDE-101\$9r:DE-101\$bger\$d1210\$erda  
=042 \\\$agnd1  
=043 \\\$cXA-DE-SN  
=075 \\\$bg\$2gndgen  
=075 \\\$bgik\$2gndspec  
=079 \\\$ag\$qs\$qh\$qq\$uw\$uz\$uv\$uo  
=083 04\$z2\$a432161\$9d:3\$9t:2010-06-07\$222/ger  
=089 04\$z2\$a4321613\$9t:2007-01-01\$9g:2010-06-07\$222/ger  
=151 \\\$aFreiberg  
=410 2\\\$aFreiberg\$bOberbürgermeister\$4spio\$4https://d-nb.info/standards/elementset/gnd#variantName\$wr\$SiSpitzenorgan\$eSpitzenorgan  
=410 2\\\$aOberbürgermeister\$gFreiberg\$4spio\$4https://d-nb.info/standards/elementset/gnd#variantName\$wr\$SiSpitzenorgan\$eSpitzenorgan  
=410 2\\\$aFreiberg\$bAmt des Oberbürgermeisters, Städtepartnerschaften\$4spio\$4https://d-nb.info/standards/elementset/gnd#variantName\$wr\$SiSpitzenorgan\$eSpitzenorgan  
=410 2\\\$aAmt des Oberbürgermeisters, Städtepartnerschaften\$gFreiberg\$4spio\$4https://d-nb.info/standards/elementset/gnd#variantName\$wr\$SiSpitzenorgan\$eSpitzenorgan  
=410 2\\\$aFreiberg\$bAbteilung Jugend, Sport und Naherholung\$4spio\$4https://d-nb.info/standards/elementset/gnd#variantName\$wr\$SiSpitzenorgan\$eSpitzenorgan

Gesamt 314.492  
Mit Koordinaten 58.758

## NER

- *Named Entity Recognition (ner) is the information extraction task of identifying and classifying mentions of people, organisations, locations and other named entities (nes) within text. [2]*

= ein Teilgebiet von NLP (Natural Language Processing)

- Grundlage: trainierte Daten

- Supervised Learning: Annotation von Objekten

```

144 Altea|NE|I-LOC hatte|VAFIN|O bei|APPR|O einer|ART|O Fläche|NN|O von|APPR|O 34,4|CARD|O km²|NN|O am|APPRART|O 1.|ADJA|O Januar|NN|O
2009|CARD|O 23.780|CARD|O Einwohner|NN|O .|$.|O
145 In|APPR|O Altea|NE|I-LOC befindet|VVFIN|O sich|PRF|O die|ART|O Fakultät|NN|O für|APPR|O Schöne|ADJA|O Künste|NN|O der|ART|O
Universität|NN|I-LOC Miguel|NE|I-LOC Hernández|NE|I-LOC Elche|NN|I-LOC .|$.|O
146 Die|ART|O Gründung|NN|O erfolgte|VVFIN|O durch|APPR|O Iberer|NN|O und|KON|O Römer|NE|I-LOC .|$.|O
147
148
149 Hermann|NE|I-PER Jüngken|NN|I-PER war|VAFIN|O ein|ART|O Rittergutsbesitzer|NN|O und|KON|O Reichstagsabgeordneter|NN|O .|$.|O
150 Jüngken|NN|I-PER war|VAFIN|O 1859|CARD|O bis|KON|O 1876|CARD|O Rittergutsbesitzer|NN|O in|APPR|O Reinsdorf|NN|I-LOC bei|APPR|O
Artern|NN|I-LOC .|$.|O
151
152 Auch|ADV|O Graf|NE|O Ludwig|NE|I-PER Heinrich|NE|I-PER von|APPR|I-PER Nassau-Dillenburg|NN|I-PER hatte|VAFIN|O sich|PRF|O um|APPR|O
ihre|PPOSAT|O Hand|NN|O bemüht|VVPP|O ,|$,|O was|PRELS|O zu|APPR|O Streitigkeiten|NN|O führte|VVFIN|O .|$.|O
153

```

# displaCy Named Entity Visualizer

Humboldt studierte in Freiberg bevor er u.a. nach Venezuela, Kuba, Kolumbien, Ecuador, Peru und Mexico reiste und schließlich 1807 nach Paris umsiedelte.



## Entity labels (select all)



PER



ORG



LOC



MISC

## Model ?

German - de\_core\_news\_sm (v2.3.0)

[3]

Humboldt **PER** studierte in Freiberg **LOC** bevor er u.a. nach Venezuela **LOC**, Kuba **LOC**, Kolumbien **LOC**, Ecuador **LOC**, Peru **LOC** und Mexico **LOC** reiste und schließlich 1807 nach Paris **LOC** umsiedelte.



## Accuracy Evaluation ⌵



TAG_ACC	Part-of-speech tags (fine grained tags, Token.tag)	0.98
ENTS_P	Named entities (precision)	0.86
ENTS_R	Named entities (recall)	0.85
ENTS_F	Named entities (F-score)	0.85
SENTS_P	Sentence segmentation (precision)	0.95
SENTS_R	Sentence segmentation (recall)	0.96
SENTS_F	Sentence segmentation (F-score)	0.95
TOKEN_ACC	Tokenization	1.00
POS_ACC	Part-of-speech tags (coarse grained tags, Token.pos)	0.98
MORPH_ACC		0.92
LEMMA_ACC		0.73
DEP_UAS	Unlabelled dependencies	0.93
DEP_LAS	Labelled dependencies	0.91

[4]

Akt. Highscore: LUKE = 94,3% (<https://github.com/studio-ousia/luke>)

## Mordecai

- Ermittelt aus unstrukturiertem Text Geo-Entities
- Nutzt spaCy (NER)
  - Textkorpus Trainingset auf Basis von Wikipedia-Artikeln [5]
  - .. und TIGER Korpus (Uni Stuttgart, Institut für Maschinelle Sprachverarbeitung, [6])
  - Annotation mittels prodigy [7]
  - KNN implementiert in Keras (Open Source DL-Bibliothek)
- Ermittelte Geo-Entities werden an Geonames-Gazetteer geschickt und Koordinaten ausgegeben
  - The GeoNames gazetteer was chosen for GeoTxt because of its extensive coverage, quality, inclusion of metadata (such as alternate names and geographic hierarchical information), and frequent updates (Acheson, Sabbata, & Purves, 2017).

## Toponym Resolution

- Über Heuristiken: population, area, or geographic-level prominence
- Co-occurrences
  - spatial minimality: spatial proximity (i.e. assuming that toponyms in a document are likely to constitute a “spatial cluster”) and therefore toponym candidates for co-occurring place names that minimize the average distance between all toponym predictions are prioritized over distant candidates
  - spatial hierarchy (country, state, county, township, populated place) or on names that belong to the same subtree of such hierarchy (counties that are both located within a particular state) [2]

Selektieren

Sammeln

Identifizieren

Disambiguieren

Alle GND-IDs ohne 034  
DNB-IDs aus 024

Ländercode 043

Ortsnamen und -varianten 151, 451?

Alle im k10+ verknüpften Objekte plus Katalogisate

Named Entity Recognition (NER)

Toponym Resolution

Geoparsing  
(zB mordecai)

## Vorgehen

- Geoparser-Software evaluieren [8]
- Korrektheit der vorgefertigten Modelle überprüfen
  - EN/DE-cores
  - Ggf. Anpassungen vornehmen
- => Intellektuell vergebenen GnD-Geo-Entities mit mordecai-Output vergleichen
- Verarbeitung mit unterschiedlich komplexen Heuristiken auswerten
  - Katalogisat
  - TOC, Abstract, Volltext
  - z.B. über SRU-K10+
    - etwa 70.000 Zeichen Zusatzinformationen

## Aktuelles Setting

- Docker Desktop mit Elasticsearch 5.5.2
- Python 3.9.4
- Tensorflow: 2.5.0
- spaCy: 3.0.6
  - en- und de-cores: 3.0.0
- Mainboard: MSI B450M PRO-VDH Plus, AMD B450
- Prozessor: AMD Ryzen 5 2600X 6x 4.2GHz
- Arbeitsspeicher: 16GB DDR4-RAM PC-3000 (2x 8GB)
- Grafikkarte: Nvidia GeForce RTX2060 6GB, Palit Gaming Pro OC
  - 1920 CUDA-Cores
- **Verarbeitung: knapp 7h**

## Beobachtungen

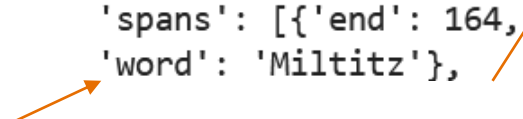
- Verarbeitung von Heuristiken  
mitunter erschwert durch Verwendung  
von Abkürzungen -> nur Titel

```
{'country_conf': 0,  
  'country_predicted': '',  
  'spans': [{ 'end': 245, 'start': 235 }],  
  'word': '2.1991,Jan'},  
{'country_conf': 0.92213047,  
  'country_predicted': 'TUR',  
  'geo': { 'admin1': 'Istanbul',  
          'country_code3': 'TUR',  
          'feature_class': 'S',  
          'feature_code': 'HTL',  
          'geonameid': '9884822',  
          'lat': '41.001',  
          'lon': '28.80273',  
          'place_name': 'Radisson Blu Conference & Airp'},  
  'spans': [{ 'end': 292, 'start': 288 }],  
  'word': 'Körp'},
```

## Beobachtungen

- Verarbeitung von Heuristiken mitunter erschwert durch Verwendung von Abkürzungen -> nur Titel
- Ausgabe von Ortsteilen / -bereichen

```
[{'country_conf': 0.9048774,  
  'country_predicted': 'DEU',  
  'geo': {'admin1': 'Saxony',  
          'country_code3': 'DEU',  
          'feature_class': 'S',  
          'feature_code': 'RSTN',  
          'geonameid': '2871073',  
          'lat': '51.32556',  
          'lon': '12.25611',  
          'place_name': 'Bahnhof Miltitz'},  
  'spans': [{'end': 164, 'start': 157}],  
  'word': 'Miltitz'},
```

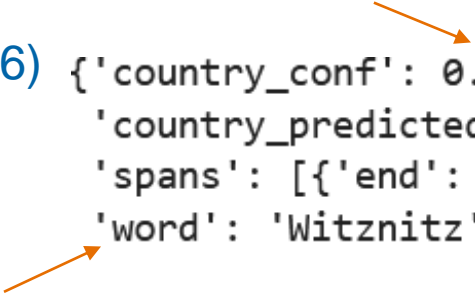
Two orange arrows point to the 'start' and 'end' values in the 'spans' array. One arrow points to 'start': 157 and the other points to 'end': 164.



## Beobachtungen

- Verarbeitung von Heuristiken mitunter erschwert durch Verwendung von Abkürzungen -> nur Titel
- Ausgabe von Ortsteilen / -bereichen
- Schwellwert anpassen (default: 0,6)

```
{'country_conf': 0.57147366,  
'country_predicted': 'DEU',  
'spans': [{'end': 8, 'start': 0}],  
'word': 'Witznitz'},
```



# Beobachtungen

- 27 / 314 Matches

dnb_id	coordinates	placename	mordecai-conf	mordecai-placer	mordecai-latlon
<a href="http://d-nb.info/gnd/126317-1">http://d-nb.info/gnd/126317-1</a>	E 011 36 01, E 011 36 01, N 050 55 57, N 050 55 57	Wenigenjena	0.9052457	Wenigenjena	<a href="https://www.latlong.net/c/?lat=50.93225&amp;long=11.59993">https://www.latlong.net/c/?lat=50.93225&amp;long=11.59993</a>
<a href="http://d-nb.info/gnd/2019606-4">http://d-nb.info/gnd/2019606-4</a>	E 012 15 17, E 012 15 17, N 051 19 28, N 051 19 28	Miltitz	0.8359622	Bahnhof Miltitz	<a href="https://www.latlong.net/c/?lat=51.32556&amp;long=12.25611">https://www.latlong.net/c/?lat=51.32556&amp;long=12.25611</a>
<a href="http://d-nb.info/gnd/3016944-6">http://d-nb.info/gnd/3016944-6</a>	E 011 19 09, E 011 19 09, N 051 08 37, N 051 08 37	Ellersleben	0.9052457	Ellersleben	<a href="https://www.latlong.net/c/?lat=51.15&amp;long=11.31667">https://www.latlong.net/c/?lat=51.15&amp;long=11.31667</a>
<a href="http://d-nb.info/gnd/4006424-4">http://d-nb.info/gnd/4006424-4</a>	E 009 47 00, E 009 47 00, N 048 06 00, N 048 06 00	Biberach an der Riß	0.8359622	Biberach an der Riß	<a href="https://www.latlong.net/c/?lat=48.08942&amp;long=9.79942">https://www.latlong.net/c/?lat=48.08942&amp;long=9.79942</a>
<a href="http://d-nb.info/gnd/4050796-8">http://d-nb.info/gnd/4050796-8</a>	E 011 42 24, E 011 42 24, N 051 06 32, N 051 06 32	Rudelsburg	0.8359622	Rummelsburg	<a href="https://www.latlong.net/c/?lat=52.50146&amp;long=13.4934">https://www.latlong.net/c/?lat=52.50146&amp;long=13.4934</a>
<a href="http://d-nb.info/gnd/4064419-4">http://d-nb.info/gnd/4064419-4</a>	E 008 16 09, E 008 16 09, N 047 37 53, N 047 37 53	Waldshut	0.8359622	Landkreis Waldshut	<a href="https://www.latlong.net/c/?lat=47.70556&amp;long=8.255">https://www.latlong.net/c/?lat=47.70556&amp;long=8.255</a>
<a href="http://d-nb.info/gnd/4104598-1">http://d-nb.info/gnd/4104598-1</a>	E 009 08 45, E 009 08 45, N 049 18 50, N 049 18 50	Burg Hornberg	0.8359622	Neckarzimmern Burg Hornberg	<a href="https://www.latlong.net/c/?lat=49.3139&amp;long=9.14583">https://www.latlong.net/c/?lat=49.3139&amp;long=9.14583</a>



## Ausblick

- Transfer Learning
- BERT (z.B. German Bert Model)
- Heuristiken (Ländercode)
- Fachdatenbank (Wikidata, etc.)

Witznitz **LOC** , zwei. Ortsfamilienbuch Witznitz **PER** 1495-1899. Witznitz II. **PER** zw  
 Sanierungsrahmenplan Tagebau Witznitz **PER** : vom  
 Sächsischen Staatsministerium des Innern am 03.02.2000 **ORG** genehmigt; Eintritt de  
 § 9 (2) SächsLPIG **MISC** am 09.09.2000. Hydrogeochemische Untersuchungen und Be  
 exothermen Reaktion in Sedimenten **LOC** im Tagebau Witznitz-Sachsen **LOC** : ein Ve  
 von Eisendisulfidverwitterungsprozessen **PER** in Kippen **LOC** . Konzepte und Strategie  
 Nachnutzung der Brikettfabrik Witznitz **ORG** . Die Hainer Sande im Tagebau Witznitz: E  
 Aufschlußdokumentation stillgelegter Braunkohlentagebaue **PER** in Sachsen **LOC** . I  
 Dorfes Witznitz **LOC** .

<https://explosion.ai/demos/displacy-ent>

Witznitz. **LOC** zwei. Ortsfamilienbuch Witznitz **LOC**  
 1495-1899. Witznitz **LOC** II. zwei. Braunkohlenplan als  
 Sanierungsrahmenplan Tagebau **LOC** Witznitz: **LOC**  
 vom Sächsischen **ORG** Staatsministerium des Innern am  
 03.02.2000 genehmigt; Eintritt der Verbindlichkeit gemäß § 9  
 (2) SächsLPIG **LOC** am 09.09.2000. Hydrogeochemische  
 Untersuchungen und Bestimmung der exothermen Reaktion in  
 Sedimenten im Tagebau Witznitz-Sachsen: **LOC** ein  
 Versuch der Erkennung von

<https://demos.deepset.ai/ner>



**Vielen Dank für Ihre Aufmerksamkeit!**

[oliver.loewe@ub.tu-freiberg.de](mailto:oliver.loewe@ub.tu-freiberg.de)

Fachinformationsdienst Montan

<https://montanportal.info>

Universitätsbibliothek Freiberg

<https://tu-freiberg.de/ub>

## Quellen

- [1] [https://grantmckenzie.com/academics/McKenzie\\_2016\\_EKAW.pdf](https://grantmckenzie.com/academics/McKenzie_2016_EKAW.pdf)
- [2] <https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12510>
- [3] <https://explosion.ai/demos/displacy-ent>
- [4] <https://spacy.io/models/de>
- [5] [https://figshare.com/articles/dataset/Learning\\_multilingual\\_named\\_entity\\_recognition\\_from\\_Wikipedia/5462500](https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500)
- [6] <https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger>
- [7] <https://prodigy.io>
- [8] <https://github.com/geoai-lab/EUPEG>