



DWDS – Architecture and Workflow

Blockseminar e-lexicography

Day 2, 13.15 – 14.45

BBAW, 2019-02-15

- I. Workflow of DWDS (part 1)
 - a. DWDS: Goals, resources, services + platform
 - b. DWDS: aggregated dictionary compilation
 - c. DWS+workflow: „a bird's eye view“
 - d. Issue management system
 - e. On-the-fly generation of dictionary entries
- II. DWS: a closer look (part 2) -> Slides Axel Herold



- 1a. project goals

compile a comprehensive corpus based dictionary
of contemporary German

- dictionary + corpus examples + word statistics
- dictionary is based on legacy dictionaries

provide knowledge about historical German
(legacy dictionaries + historical German)

(A) 465 000 dictionary entries

→ DWDS-Sync (i.e. WDG (1961-77) + Duden 1999), +
DWDS-Hist (Grimm, Etymological Dictionary),
Thesauri (OpenThesaurus, GermaNet)

(B) Corpora (13 billion tokens)

→ reference corpora (17th- 20th c.), national daily +
weekly newspapers, CMC-corpora

(C) Word statistics

→ time lines, word sketches



Leiter



Startseite / Wortinformation zu „Leiter“

Leiter¹

Leiter²

Leiter, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Leiter** · Nominativ Plural: **Leitern**

Aussprache

Worttrennung Lei-ter (computergeneriert)

Wortbildung mit ›Leiter‹ als Erstglied: [Leitersprosse](#) ... 2 weitere · mit ›Leiter‹ als Letztglied: [Aalleiter](#) ... 25 weitere

Bedeutung

Etymologie

Thesaurus

Typische Verbindungen

Verwendungsbeispiele

Bedeutung

eWDG, 1969

- Gerät aus Holz oder Leichtmetall, das aus zwei durch mehrere Sprossen verbundenen Längsstangen besteht und zum Hinaufsteigen dient

BEISPIELE:

eine hohe, niedrige, lange, ausziehbare, eiserne **Leiter**

auf eine **Leiter** klettern, steigen

die **Leiter** an einen Baum lehnen, anlegen

... 5 weitere Beispiele

▼ **bildlich**

BEISPIELE:

er stieg die **Leiter** des Erfolges, Ruhmes rasch empor

er steht auf der **Leiter** (= der obersten Sprosse der Leiter) der höchsten Macht

Worthäufigkeit

selten häufig

Frequenz / Mio Tokens



Ältere Wörterbücher

[Grimmsches Wörterbuch \(1 DWB\)](#) (2)

[Wörterbuch der deutschen Gegenwartssprache \(WDG\)](#) (2)

Korpustreffer

Referenzkorpora

[DWDS-Kernkorpus \(1900–1999\)](#) (8310)

[DWDS-Kernkorpus 21 \(2000–2010\)](#) (485)

[Deutsches Textarchiv \(1473–1927\)](#) (4359)

Zeitungskorpora

[Berliner Zeitung \(1994–2005\)](#) (17352)

[Tagesspiegel \(1996–2005\)](#) (10284)

[Die ZEIT \(1946–2016\)](#) (27374)

Spezialkorpora

[Referenz- und Zeitungskorpora](#)

www.dwds.de/wb/Leiter ; engl. ladder (¹Leiter) oder leader (²Leiter)

Etymologie

Etymologisches Wörterbuch (Wolfgang Pfeifer)

Leiter · Leiterwagen

Leiter f. Gerät aus zwei durch Sprossen verbundenen Längsstangen (Holmen) zum Hinaufsteigen, ahd. (*h*)*leitara* (9. Jh.), mhd. *leiter*, *leitere*, mnd. *ladder*, mnl. *lāder*, nl. *leer* und (mit Vokalwechsel) mnl. *ladder(e)*, *lāder(e)*, nl. *ladder*, aengl. *hlæd(d)er*, engl. *ladders* sind mit dem Suffix für Gerätebezeichnungen ie. *-tro-* bzw. hier mit der fem. Suffixvariante *-trā* gebildet zu der unter ↗ ... **Mehr**

leiten · Leiter · Leiterin · Leitung · anleiten · Anleitung · einleiten · Einleitung · geleiten · Geleit · verleiten · Leitartikel · Leitfaden · Leithammel · Leitstern

leiten Vb. 'führen, an der Spitze stehen und den Verlauf bestimmen, befehligen, in eine bestimmte Richtung lenken, Elektrizität, Wärme, Schall durch einen Stoff weiterführen', ahd. (8. Jh.), mhd. *leiten* 'geleiten, lenken, führen', asächs. *lēdian*, mnd. *lēden*, mnl. *leiden*, *lēden*, nl. *leiden*, aengl. *lædan*, engl. *to lead*, anord. *leiða*, schwed. *leda* ist ein Kausativum im Sinne von 'Fortbewegung ... **Mehr**

Thesaurus

www.openthesaurus.de (05/2017)

► Synonymgruppe

↗**Boss** · ↗**Chef** · ↗**Dienstherr** · Dienstvorgesetzter · ↗**Geschäftsherr** · ↗**Superior** · Vorgesetzter • ↗**(der) Alte** ugs., salopp · Chef von't Janze ugs., salopp, berlinerisch · Obermacker ugs., salopp · Obermufti ugs., fig., salopp · ↗**Prinzipal** geh., veraltet · hohes Tier ugs.

► Synonymgruppe

↗**Entscheider** · ↗**Entscheidungsträger** · ↗**Führungskraft** · ↗**Manager** • ↗**Lenker** ugs.

► Synonymgruppe

↗**Anführer** · ↗**Führer** · ↗**Hauptmann** · ↗**Häuptling** · ↗**Oberhaupt** • ↗**Kopf** ugs.

► Synonymgruppe

↗**Sprossenleiter** · Sprossenstiege • ↗**Fahrt (Bergmannsspr.)** fachspr., Jargon

Synonymgruppe

↗**Aufseher** · ↗**Intendant** · ↗**Leiter (Rundfunk-/Fernsehanstalt)**

► Synonymgruppe

↗**L** fachspr. · ↗**Phase** ugs.

Spezialkorpora

Referenz- und Zeitungskorpora
(aggregiert) (40495)

Blogs (1961)

Polytechnisches Journal (3412)

Filmuntertitel (1471)

Gesprochene Sprache (63)

DDR (600)

Typische Verbindungen

computergeneriert

DWDS-Wortprofil



Detailliertere Informationen bietet das [DWDS-Wortprofil zu >Leiter<](#).

Verwendungsbeispiele

maschinell ausgesudt aus den DWDS-Korpora

DWDS-Beispielextraktor

Zur Lösung der Probleme fällt dem Sportlichen **Leiter** aber auch nicht mehr viel ein.

Die Welt, 06.05.2003

Wie bestellt ist in der Nähe auch eine passende **Leiter** zur Hand.

Das Tennismagazin, 08.11.2001




Leiter

Leiter¹Leiter²

Leiter, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Leiter** · Nominativ Plural: **Leitern**

Aussprache 

Worttrennung Lei-ter (computergeneriert)

Wortbildung mit ›Leiter‹ als Erstglied: ↗[Leitersprosse](#) · ↗[Leiterwagen](#) · ↗[leiterartig](#) ... **weniger**

mit ›Leiter‹ als Letztglied: ↗[Aalleiter](#) · ↗[Anlegeleiter](#) · ↗[Badeleiter](#) · ↗[Bockleiter](#) · ↗[Brandleiter](#) · ↗[Bücherleiter](#) · ↗[Drehleiter](#) ·
↗[Erfolgsleiter](#) · ↗[Feuerleiter](#) · ↗[Feuerwehrleiter](#) · ↗[Froschleiter](#) · ↗[Hakenleiter](#) · ↗[Himmelsleiter](#) · ↗[Hühnerleiter](#) ·
↗[Jakobsleiter](#) · ↗[Karriereleiter](#) · ↗[Rangleiter](#) · ↗[Sprossenleiter](#) · ↗[Stehleiter](#) · ↗[Steigeleiter](#) · ↗[Strickleiter](#) · ↗[Stufenleiter](#) ·
↗[Sturmleiter](#) · ↗[Tonleiter](#) · ↗[Treppenleiter](#) · ↗[Trittleiter](#) ... **weniger**

Bedeutung

Etymologie

Thesaurus

Typische Verbindungen

Verwendungsbeispiele

Bedeutung

eWDG, 1969

- ✓ Gerät aus Holz oder Leichtmetall, das aus zwei durch mehrere Sprossen verbundenen

Leiter, der

Grammatik Substantiv (Maskulinum) · Genitiv Singular: **Leiters** · Nominativ Plural: **Leiter**

Aussprache 

Worttrennung Lei-ter (computergeneriert)

Wortzerlegung [Leiten](#) [-er](#)

Wortbildung mit ›Leiter‹ als Erstglied: [Leiterbahn](#) ... **1 weitere** · mit ›Leiter‹ als Letztglied: [Absatzleiter](#) · [Abstimmungsleiter](#) · [Abteilungsleiter](#) · [Amtsleiter](#) · [Anstaltsleiter](#) · [Arbeitsgruppenleiter](#) · [Aufnahmeleiter](#) · [Ausstattungsleiter](#) · [Bauleiter](#) · [Behördenleiter](#) · [Bereichsleiter](#) · [Betriebsleiter](#) · [Bezirksleiter](#) · [Brigadeleiter](#) · [Bühnenleiter](#) · [Bündelleiter](#) · [Büroleiter](#) · [Chorleiter](#) · [Delegationsleiter](#) · [Dienststellenleiter](#) · [Diskussionsleiter](#) · [Distriktleiter](#) · [Distriktsleiter](#) · [Eileiter](#) · [Einsatzleiter](#) · [Elektrizitätsleiter](#) · [Expeditionsleiter](#) · [Fachbereichsleiter](#) · [Fachgruppenleiter](#) · [Fahrdienstleiter](#) · [Festivalleiter](#) · [Filialleiter](#) · [Forschungsgruppenleiter](#) · [Gauleiter](#) · [Gebietsleiter](#) · [Geschäftsstellenleiter](#) · [Gesprächsleiter](#) · [Gewerkschaftsgruppenleiter](#) · [Grundwasserleiter](#) · [Gruppenleiter](#) · [Halbleiter](#) · [Harnleiter](#) · [Hauptabteilungsleiter](#) · [Hauptamtsleiter](#) · [Heimleiter](#) · [Institutsleiter](#) · [Interimsleiter](#) · [Kaderleiter](#) · [Klassenleiter](#) · [Kreisleiter](#) · [Kulturleiter](#) · [Kursleiter](#) · [Küchenleiter](#) · [Lagerleiter](#) · [Mannschaftsleiter](#) · [Marketingleiter](#) · [Mitleiter](#) · [Museumsleiter](#) · [Nichtleiter](#) · [Niederlassungsleiter](#) · [Nullleiter](#) · [Objektleiter](#) · [Ortsgruppenleiter](#) · [Personalleiter](#) · [Pionierleiter](#) · [Planungsleiter](#) · [Politleiter](#) · [Produktionsleiter](#) · [Programmleiter](#) · [Projektleiter](#) · [Redaktionsleiter](#) · [Referatsleiter](#) · [Regionalleiter](#) · [Reisegruppenleiter](#) · [Reiseleiter](#) · [Rennleiter](#) · [Ressortleiter](#) · [Sachgebietsleiter](#) · [Samenleiter](#) · [Schichtleiter](#) · [Schriftleiter](#) · [Schulleiter](#) · [Schulungsleiter](#) · [Seminarleiter](#) · [Sendeleiter](#) · [Spielleiter](#) · [Stromleiter](#) · [Studiengruppenleiter](#) · [Studienleiter](#) · [Supraleiter](#) · [Teamleiter](#) · [Theaterleiter](#) · [Transportleiter](#) · [Veranstaltungsleiter](#) · [Verkaufsleiter](#) · [Verkaufsstellenleiter](#) · [Verlagsleiter](#) · [Versammlungsleiter](#) · [Versandleiter](#) · [Versuchsleiter](#) · [Verwaltungsleiter](#) · [Wahlkampfleiter](#) · [Wahlleiter](#) · [Werbeleiter](#) · [Werkleiter](#) · [Werksleiter](#) · [Wärmeleiter](#) · [Zuchtleiter](#) · [Übungsleiter](#) ... **weniger**





Aggregation of dico info

- Form info: compiled by DWDS
- Sense info: aggregated from:
 - eWDG: 120,000 entries (CR – BBAW)
 - Duden: 70,000 entries (licensed from BI)
 - Wahrig: 5,000 entries (licensed from Wahrig)
 - ~1,000 (~3,000) entries published (compiled) by DWDS
- Etymological info: Pfeifer
- Legacy Dictionary info : WDG, Grimm (¹DWB, ²DWB, Sanders)



- Mixed sources in an entry
- Source description of an entry:
 - WDG | Duden | Wahrig | Duden_DWDS | Wahrig_DWDS
 - Duden_DWDS stands for: entry+def by Duden, examples+collocations by DWDS (e.g. Tonsprache (tonal language))
- Mixed sources on lower levels, e.g.
 - 2 senses by WDG, 2 senses by DWDS (e.g. Maus)
 - Entry by WDG, updated examples by DWDS

Ic. DWS: Bird's eye view

- DWDS is used as corpus search platform by lexicographers
- dictionary schema
- oXygen + exist
- Stages in the compilation of entries: draft, red_1, red_2, red_f
- DWDS_schema: can be converted to tei (but not lossless)

Id. Issue management system

- Users can report wrong&missing information:
 - Entries, senses, form....
- User front end: web formular
- Back end: mantis
- 34,000 issues reported (by ~30 persons)

le. On-the-fly generation

1. problem statement: out of dictionary queries
2. Automatic Morphological Segmentation
 - Automatic Morphological Analysis (MA)
 - Mapping MA to dictionary headwords
3. Components of dynamically generated entries
4. Fallback Mechanisms
5. Results and Discussion



Legacy dictionaries

- WDG: 120,000 headwords
- Duden: 200,000 headwords
- Grimm: ~335,000 headwords (estimated)
- German vocabulary: unknown and potentially unlimited due to rich compounding rules
 - ~ 5 million lemmas for a corpus of 1 billion tokens of German (e.g. Klein 2013)
 - ~ 14.95 million lemmas (recognized by automatic morphological analysis) in a 4 billion corpus (DWDS German Newspaper Corpus)
 - Follows Heaps' law: $V = k * n^{\text{beta}}$

Problem statement

- How many headwords are lexicalized?
- How many headwords are „out of legacy dictionary“ queries?
- What information should be provided for those queries?

rich word formation in German: two examples

derivation:

$(((\text{voll}_P \text{streck}_V) \text{bars}) \text{keits})_N$ (engl. enforceability)

composition:

$(((\text{voll}_P \text{streck}_V) \text{bars}) \text{keits})_N \text{ s}_{\text{LINK}} ((\text{er}_P \text{klär}_V)_V \text{ung}_S)_N$

Mapping MA to dictionary entries

FSM morphological analyzers for German:

- a. Gertwol (Haapalainen and Majorin, 2003)
- b. SMOR (Schmid, 2004)
- c. TAGH (Geyken & Hanneforth, 2006)

ad a. Gertwol no freely reuse

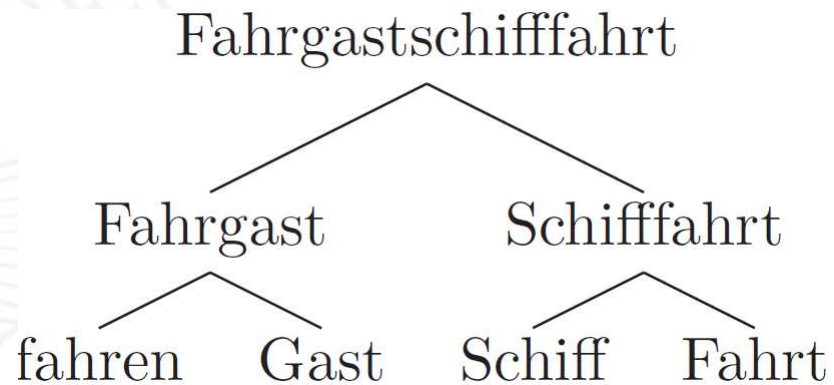
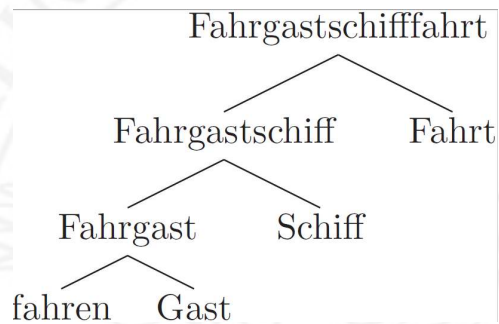
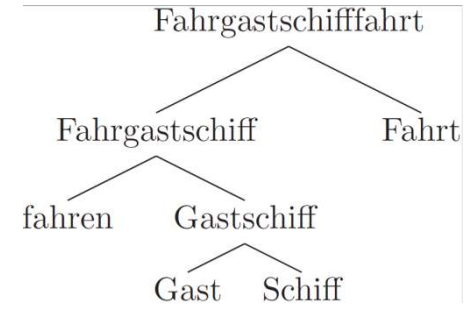
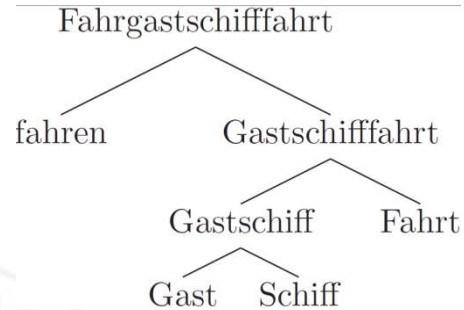
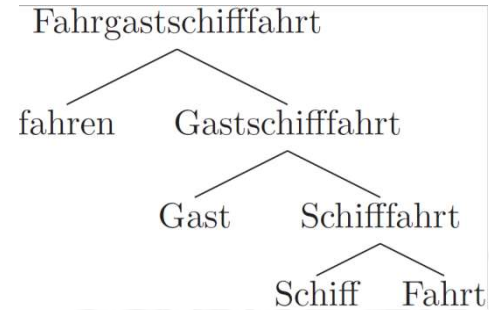
b+c. SMOR allows for an „atomic“ segmentation
whereas TAGH lexicon regroups morphemes to
larger units

Example

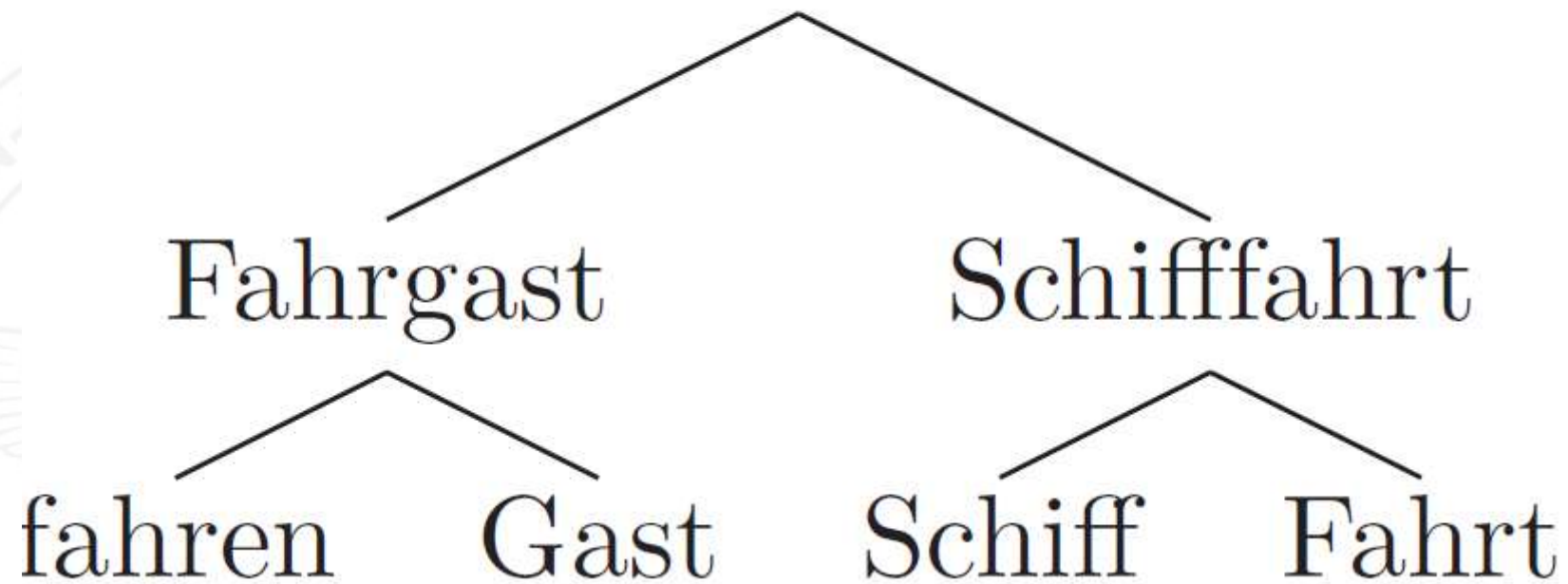


Fahrgastschiffahrt - passenger shipping

MA of „Fahrgastschiffahrt“



Fahrgastschiffahrt



Components of „on-the-fly-articles

- Form: spelling, grammar, word formation, word frequency
- Thesaurus (if available from OpenThesaurus)
- DWDS – Good Example Extractor
- DWDS – collocation Extractor (word profile)
- Corpus hits + inflected forms

Fahrgastschiffahrt



[Startseite](#) / Wortinformation zu „Fahrgastschiffahrt“

„Fahrgastschiffahrt“ ist nicht in unseren gegenwartssprachlichen lexikalischen Quellen vorhanden. Folgende Informationen konnten automatisch ermittelt werden:

Fahrgastschiffahrt, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Fahrgastschiffahrt** · Nominativ Plural: **Fahrgastschiffahrten** (computergeneriert)

Worttrennung Fahr-gast-schiff-fahrt

Wortzerlegung ↗Fahrgast ↗Schiffahrt

Verwendungsbeispiele

DWDS-Beispielextraktor

maschinell ausgedacht aus den DWDS-Korpora

Doch nicht nur die **Fahrgastschiffahrt** sieht in der Regelung Probleme.

Die Welt, 31.03.2000

Die Fachleute aus zehn Ländern erörtern Möglichkeiten zur Verbesserung von Abläufen in der **Fahrgastschiffahrt**.

Die Welt, 30.04.2004

Wegen der niedrigen Oder gibt's keine **Fahrgastschiffahrt** im Raum Frankfurt mehr.

Bild, 06.08.2004

Die Müritz hat 117 Quadratkilometer Wasserfläche und gehört zu den Mecklenburger Oberseen, die inzwischen generell für die **Fahrgastschiffahrt** gesperrt sind.

Die Zeit, 24.01.2013 (online)

Dies hat kürzlich der Verband der **Fahrgastschiffahrt** beschlossen, dem bayernweit 24 Reedereien angehören.

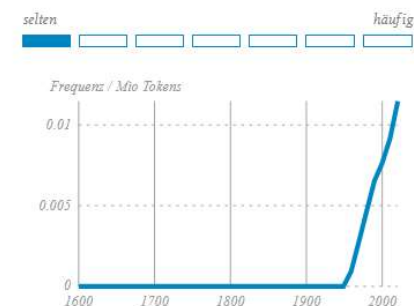
Süddeutsche Zeitung, 01.04.2000

Zitationshilfe

„Fahrgastschiffahrt“, bereitgestellt durch das Digitale Wörterbuch der deutschen Sprache, <<https://www.dwds.de/wb/Fahrgastschiffahrt>>, abgerufen am 15.09.2017.

Weitere Informationen ...

Worthäufigkeit



Ältere Wörterbücher

- [Grimmsches Wörterbuch \(¹DWB\)](#) (0)
- [Wörterbuch der deutschen Gegenwartssprache \(WDG\)](#) (0)

Korpustreffer

Referenzkorpora

- [DWDS-Kernkorpus \(1900–1999\)](#) (0)
- [DWDS-Kernkorpus 21 \(2000–2010\)](#) (0)
- [Deutsches Textarchiv \(1473–1927\)](#) (0)

Zeitungskorpora

- [Berliner Zeitung \(1994–2005\)](#) (8)
- [Tagesspiegel \(1996–2005\)](#) (18)
- [Die ZEIT \(1946–2016\)](#) (4)

Spezialkorpora

- [Referenz- und Zeitungskorpora \(aggregiert\)](#) (4)

Fahrgastschiffahrt



„Fahrgastschiffahrt“ ist nicht in unseren gegenwartssprachlichen lexikalischen Quellen vorhanden. Folgende Informationen konnten automatisch ermittelt werden:

Fahrgastschiffahrt, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Fahrgastschiffahrt** · Nominativ Plural: **Fahrgastschiffahrten** (computergeneriert)

Worttrennung Fahr-gast-schiff-fahrt

Wortzerlegung ↗Fahrgast ↗Schiffahrt

Verwendungsbeispiele

DWDS-Beispielextraktor

maschinell ausgedacht aus den DWDS-Korpora

Doch nicht nur die **Fahrgastschiffahrt** sieht in der Regelung Probleme.

Die Welt, 31.03.2000

Die Fachleute aus zehn Ländern erörtern Möglichkeiten zur Verbesserung von Abläufen in der **Fahrgastschiffahrt**.

Die Welt, 30.04.2004

Wegen der niedrigen Oder gibt's keine **Fahrgastschiffahrt** im Raum Frankfurt mehr.

Bild. 06.08.2004



Oktoberfest: tent of Hacker-Pschorr © Imago



toilet key of the canteen of the Oktoberfest

Oktoberfestkantinentoilettenschlüssel



„Oktoberfestkantinentoilettenschlüssel“ ist nicht in unseren gegenwartssprachlichen lexikalischen Quellen vorhanden. Folgende Informationen konnten automatisch ermittelt werden:

Oktoberfestkantinentoilettenschlüssel

Oktoberfestkantinentoilettenschlüssel

Oktoberfestkantinentoilettenschlüssel, der

Grammatik Substantiv (Maskulinum) · Genitiv Singular: **Oktoberfestkantinentoilettenschlüssels** · Nominativ Plural:

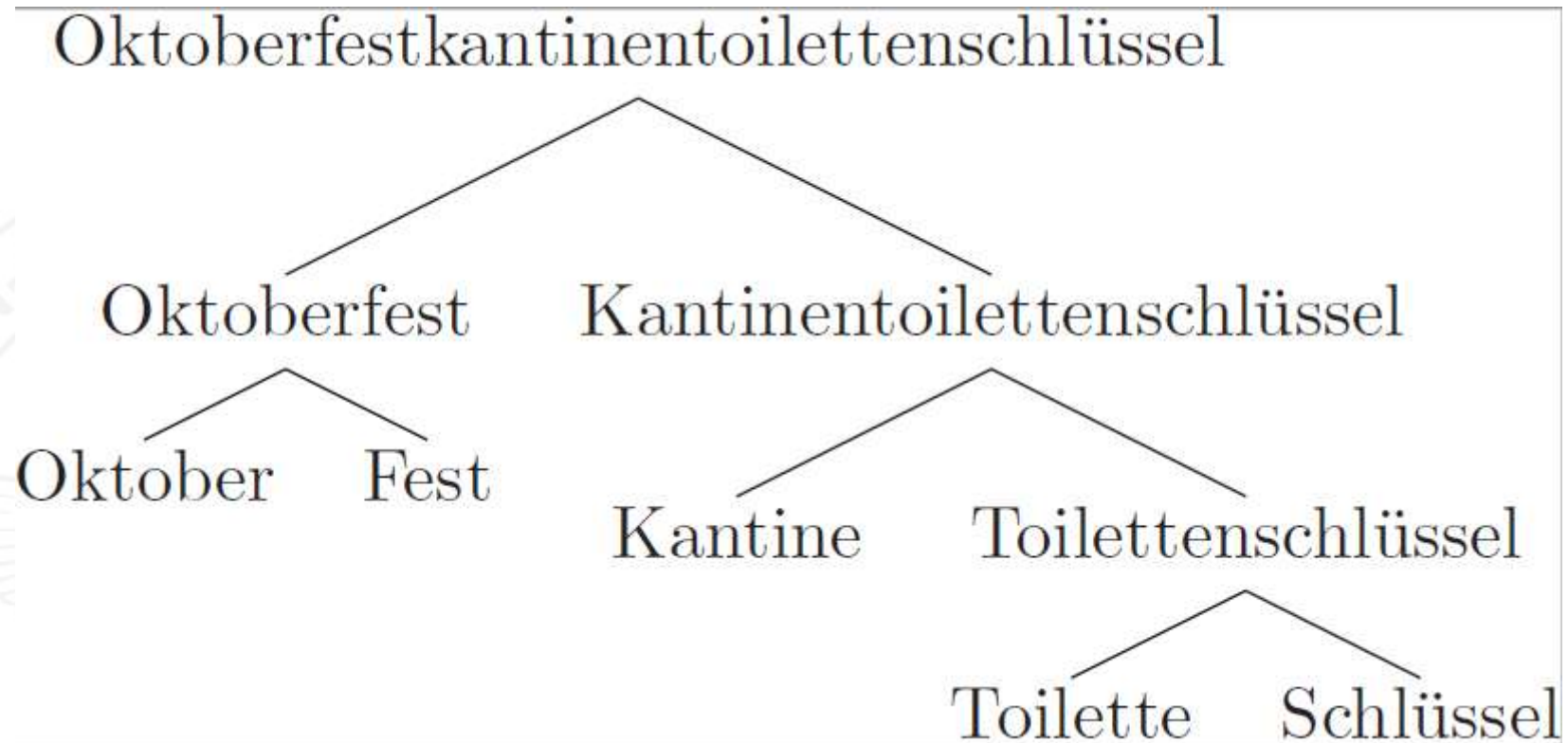
Oktoberfestkantinentoilettenschlüssel (computergeneriert)

Worttrennung Ok-to-ber-fest-kan-ti-nen-toi-let-ten-schlüs-sel

Wortzerlegung ↗Oktoberfest ↗Kantine ↗Toilette ↗Schlüssel

Zitationshilfe

„Oktoberfestkantinentoilettenschlüssel“, bereitgestellt durch das Digitale Wörterbuch der deutschen Sprache, <<https://www.dwds.de/wb/Oktoberfestkantinentoilettenschlüssel#1>>, abgerufen am 15.09.2017.



toilet key of the canteen of the Oktoberfest

Fallback mechanisms

(1) If query string q can be analyzed by SMOR
then

(a) if q = inflected form of dict. headword h
then redirect query to h

(b) else if SMOR provides segmentation
 $q=s_1...s_n$ and $s_1...s_n$ correspond to h
then generate on the-fly-article

Fallback mechanisms

(2) If query string q cannot be analyzed by SMOR then

trigger „did you mean“ function (aim is to refer user to a close form (in terms of edit distance))

(3) else refer user to a corpus search

Results

Fallback method	% of total	% correct
(1a) Inflected input query	35 %	91 %
(1b) On the fly article	20 %	95 %
(2) „Did you mean	28 %	68 %
(3) Redirect to corpus	17 %	n/a

basis: user queries for dwds (17-04-23 to 17-05-23):

- 190,554 Total number of queries
- 33,134 not in legacy dictionaries

Ambiguities of multiple segmentations can be correctly segmented with the help of the headwords of the legacy dictionaries.

Ex.: Angsthasenpolitik (engl. politics of cowardice)

(1a) Angsthase # Politik (correct)

(1b) Angst # Hasenpolitik (wrong)

Discussion

Ambiguities of linking elements:

Ex.: Reiseabschnitt (engl. travel segment)

(2a) Reis \e # Abschnitt (wrong)

(2b) Reise # Abschnitt (correct)

Discussion

Mapping to the correct word category :

Ex.:

Grillfest (engl.)



www.shutterstock.com - 428921062

(2a) Grill\N # Fest\N (wrong)

(2b) grillen\V # Fest\N (correct)

Conclusion

Evaluation showed (1 month log file):

- 1 query out of 6: out of headword query (ohq)
- 35% of ohq correspond to an inflected form
- 20% of ohq an on-the-fly-article can be generated (95% correct)
- 28% small edit distance => „did you mean?“
- 17% no morphological analysis (MA) possible

=> MA useful extension of a German dictionary
when combined with dictionary headwords