Using an Alignment-based Lexicon for Canonicalization of Historical Text

Bryan Jurish,
Henriette Ast

Deutsches Textarchiv
Berlin-Brandenburgische Akademie der Wissenschaften
jurish@bbaw.de

Historical Corpora 2012

Goethe Universität, Frankfurt am Main, 6th-8th December, 2012

Overview

The Big Picture

- Canonicalization
- Aligned Corpus
- Alignment-based Lexicon

Nasty Surprises

- Identity Pairs
- Sanitation Engineering
- Trimmed Corpus

Experiments

- Method
- Results

Conclusion

— The Big Picture —

Canonicalization

a.k.a. (orthographic) 'standardization', 'normalization', 'modernization', . . .

The Problem

- Historical text $\not\ni$ orthographic conventions
- Conventional NLP tools $\Rightarrow$ strict orthography
  - ***Fixed lexicon*** keyed by ***orthographic form***
  - ***Extant*** lexemes only

þe	Olde	Wydgette	Shoppe
↓	↓	↓	↓
the	*old*	*widget*	*shop*

The Approach

- *Map* each word w to a unique canonical cognate $\widetilde{w}$
  - Synchronically active "extant equivalents"
  - Preserve both ***root*** and ***relevant features*** of input
- *Defer* application analysis to canonical forms

Aligned Corpus

Ground-Truth Canonicalizations

- Manually verified canonicalization pairs $(w \mapsto \widetilde{w})$
- Full sentential context

Intuitions

1. Contemporary editions $\Longrightarrow$ ***already standardized***
2. Expect mostly ***identity canonicalizations*** $(w = \widetilde{w})$

Construction (sketch)

(Jurish, Drotschmann & Ast, [forthcoming])

- ***Align*** historical text with a contemporary edition
  - maximize identity alignments
- ***Confirm*** or ***reject*** type-wise alignments
- ***Manually annotate*** only unconfirmed tokens
- 126 volumes (1780–1901), 5.6M tokens, 212k types

Alignment-based Lexicon

Basic Idea

- Deterministic type-wise mapping $\boxed{\mathsf{LEX} : \mathcal{A}^* \to \mathcal{A}^* : w \mapsto \widetilde{w}}$
- Choose ***most frequent*** modern form for each input word
  - use ***string identity*** fallback for unknown words

Expected Weaknesses

(Kempken et al., 2006; Gotscharek et al., 2009b)

- Can't handle any *ambiguity*
- Identity fallback $\rightsquigarrow$ *sparse data effects*
  - especially for productive morphological processes

Alternatives

- ID: naïve string identity baseline
- HMM: robust generative HMM canonicalizer *(Jurish, 2010c; 2012)*
- HMM+LEX: alignment-based lexicon with HMM fallback
- *. . . and more!*

— Nasty Surprises —

(and some ways to deal with them)

Nasty Surprises

Intuition (1) Violations

- Assumed: modern edition $\Longrightarrow$ strict orthography
- Implicitly accepted **identity pairs** $(w \mapsto w)$
  - *ca.* 59% types, 87% tokens identical modulo transliteration
- Not always justified by the editions used ***(oops)***

Letter Case	bruder $\mapsto$ bruder	$\neq$ Bruder	*"brother"*
	trost $\mapsto$ trost	$\neq$ Trost	*"comfort"*
Extinct Forms	ward $\mapsto$ ward	$\neq$ wurde	*"was"*
	däuchte $\mapsto$ däuchte	$\neq$ dünkte	*"seems"*
Prosodic Foot	andre $\mapsto$ andre	$\neq$ andere	*"other"*
	eignen $\mapsto$ eignen	$\neq$ eigenen	*"own"*
Dialect	kömmt $\mapsto$ kömmt	$\neq$ kommt	*"comes"*
	nich $\mapsto$ nich	$\neq$ nicht	*"not"*
Apostrophes	in's $\mapsto$ in's	$\neq$ ins	*"into the"*
	s'ist $\mapsto$ s'ist	$\neq$ es ist	*"it is"*

Sanitation Engineering

a.k.a. 'garbage disposal'

Coarse Pruning (by Region)

- Dropped 5 volumes *: verse, case, dialect*
- Dropped 204 pages in 41 volumes *: dialect, foreign material*
- 245k tokens $\sim$ 32k types $\sim$ 12k local types

Heuristic Pruning (by Type)

- Invalidated all types containing
  - apostrophes or quotation marks
  - mixture of alphabetic and non-alphabetic characters
- 16k tokens $\sim$ 9k types

The Usual Suspects (under review)

- Inconsistency with respect to *online error database*
- Unknown "modern" forms (TAGH) *(Geyken & Hanneforth, 2006)*
- 57k tokens $\sim$ 12k types marked *"suspicious"*
  - currently 55k tokens, 10k suspicious types *re-validated*

Corpus Summary

Text Resources

- Source texts: *Deutsches Textarchiv (DTA)*
  - *Belles lettres*, drama, verse, philosophy *(1780–1901)*
- Target texts: `gutenberg.org`, `zeno.org`
- 126 volumes $\sim$ 5.6M tokens $\sim$ 212k pair-types

Corpus Pruning

- Removed *all sentences* containing “suspicious” material
- 13% tokens $\sim$ 18% types

Trimmed Corpus

- 121 volumes $\sim$ 4.9M tokens $\sim$ 174k types

— Experiments —

Method

‘Prototype’ Corpus $\rightsquigarrow$ Ground-Truth Relevance

$$\mathrm{relevant}(w, \widetilde{w}) := \{(v, \widetilde{v}) : \widetilde{v} = \widetilde{w}\}$$

- Most thoroughly annotated corpus subset
- 13 volumes $\sim$ 320k tokens $\sim$ 28k types (words only)

Training Corpus $\rightsquigarrow$ Canonicalization Lexicon (LEX)

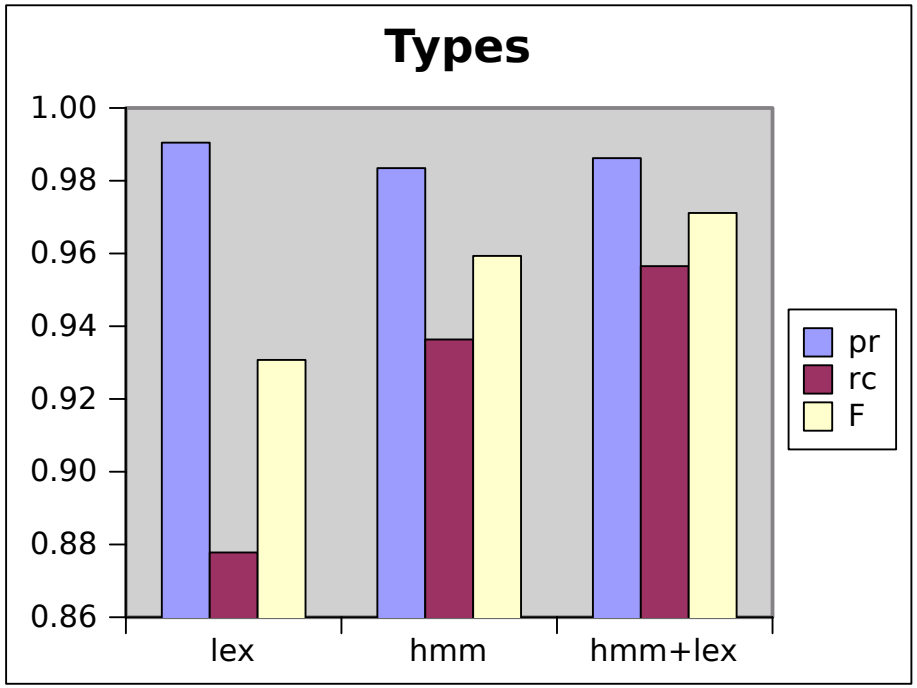
$$\mathsf{LEX}(w) = \begin{cases} \arg\max_{\widetilde{w}} f(w, \widetilde{w}) & \text{if } f(w) > 0 \\ w & \text{otherwise} \end{cases}$$

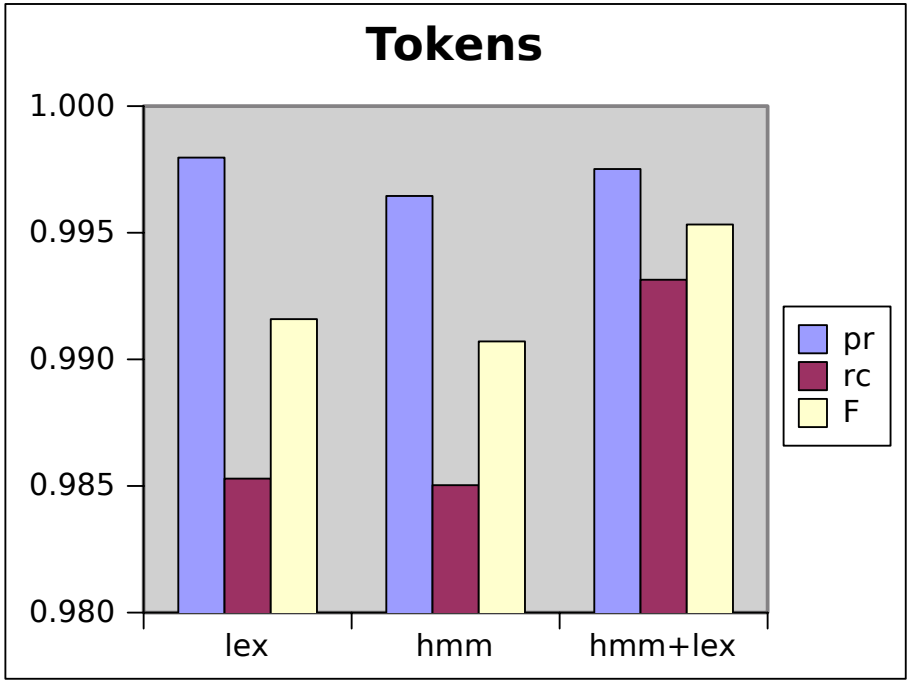
- Strictly disjoint from test corpus (by author)
- 101 volumes $\sim$ 3.5M tokens $\sim$ 158k types (words only)

Evaluation

- Simulated information retrieval $(\mathrm{pr}, \mathrm{rc}, \mathrm{F})$ *(van Rijsbergen, 1979)*
- Tested methods: ID, LEX, HMM, HMM+LEX

Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
ID	***99.1***	55.7	71.3	***99.8***	78.5	87.9
LEX	99.0	87.8	93.1	***99.8***	98.5	99.2
HMM	98.3	93.6	95.9	99.6	98.5	99.1
HMM+LEX	98.6	***95.7***	***97.1***	***99.8***	***99.3***	***99.5***

Conclusion

Aligned Corpus

- Fast bootstrapping for a canonicalization lexicon
- . . . but beware of identity mappings!

Alignment-based Canonicalization Lexicon

- Surprisingly effective on its own
  - very high precision
  - mediocre recall for unknown types (sparse data)
- Better as ‘exception’ lexicon for HMM canonicalizer
  - best overall performance
  - corpus-based and generative techniques

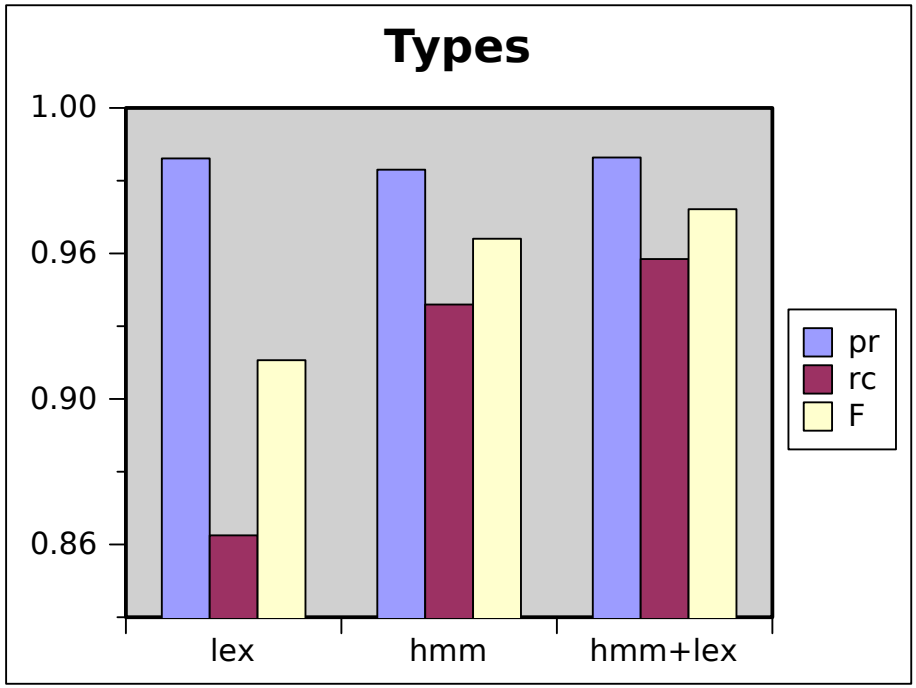
complement one another

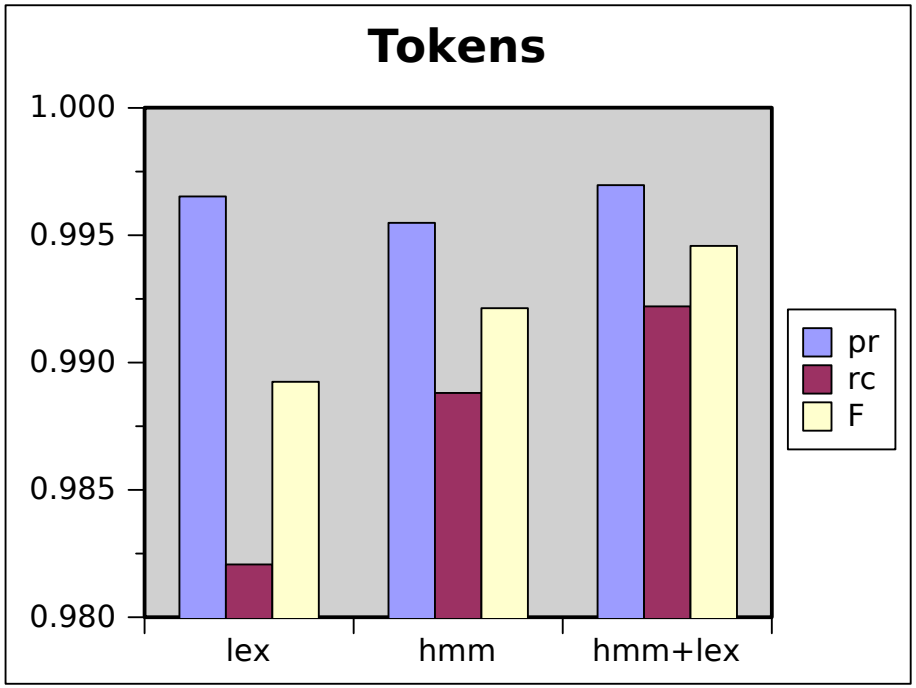
þe Olde Laste Slyde
(“The End”)

Thank you for listening!

— Addenda —

Results (pre-cleanup)





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
ID	***98.3***	57.1	72.2	***99.7***	79.1	88.2
LEX	***98.3***	85.3	91.3	***99.7***	98.2	98.9
HMM	97.9	93.2	95.5	99.5	98.9	99.2
HMM+LEX	***98.3***	***94.8***	***96.5***	***99.7***	***99.2***	***99.5***

Pruning Tool: Document List

http://kaskade.dwds.de/dta-ecp/view.perl

DTA::EvalCorpus::Prune: View

USER: moocow

SELECT: dtaid, dtadir, status, updater,updated, ntyp,ntok, pptyp_ok,pptok_ok

FROM: vdoc

WHERE:

GROUP BY:

ORDER BY: ptyp_ok asc

OFFSET: 0 LIMIT: 10 submit

Home First << Prev Next >>

	dtaid	dtadir	status	updater	updated	ntyp	ntok	pptyp_ok	pptok_ok
Edit	16211	george_algabal_1892	drop	HenrietteAst	2012-04-02 12:06:57	1178	1986	72.5 %	82.8 %
Edit	16240	george_seele_1897	drop	HenrietteAst	2012-04-02 12:27:45	2896	7002	74.4 %	85.4 %
Edit	16661	grimm_maerchen02_1815	purge	HenrietteAst	2012-04-02 13:37:23	10644	95788	75.4 %	92.0 %
Edit	16660	grimm_maerchen01_1812	purge	HenrietteAst	2012-04-02 14:03:56	11040	111223	78.1 %	93.7 %
Edit	16258	nestroy_lumpacivagabundus_1835	drop	HenrietteAst	2012-04-02 13:38:30	2869	13294	79.3 %	92.1 %
Edit	16284	kleist_krug_1811	purge	HenrietteAst	2012-04-04 15:16:50	3512	19163	88.9 %	95.6 %
Edit	17203	mueller_waldhornist_1821	purge	HenrietteAst	2012-04-04 15:51:10	3048	11917	88.9 %	95.4 %
Edit	16411	schnitzler_liebelei_1896	purge	HenrietteAst	2012-04-10 11:03:30	2134	14663	89.4 %	96.1 %
Edit	16393	goethe_iphigenie_1787	purge	HenrietteAst	2012-04-10 11:22:44	4046	17089	90.0 %	96.2 %
Edit	16181	goethe_faust01_1808	purge	HenrietteAst	2012-04-10 11:08:49	6881	34042	90.0 %	96.5 %

Home First << Prev Next >> jurish@bbaw.de

DTA::EvalCorpus::Prune 0.02 / DbCgi version 0.02

Pruning Tool: Properties

http://kaskade.dwds.de/dta-ecp/edit.perl?doc=39#tabProps

DTA::EvalCorpus::Prune: Edit (grimm_maerchen01_1812)

Home | Back | Revert | Apply

Properties | Plot | Pairs

DTADIR: grimm_maerchen01_1812 DTAID: 16660 DOC: 39
UPDATED: HenrietteAst 2012-04-02 14:03:56
TYPES: 8620 safe / 11040 total = **78.1 % safe**
TOKENS: 104163 safe / 111223 total = **93.7 % safe**
STATUS: purge (basically good data; errors can be heuristically purged)

FLAGS
- [ ] nlower
- [ ] dialogue
- [x] dialect
- [x] fm
- [ ] tokpunct
- [ ] dbclobber
- [x] blockprune
- [ ] other

COMMENTS:
blocks:
Seite 103 - 110 Dialekt
Seite 238 - 250 Dialekt
Seite 426, 429 FM
Seite 434 - 437 Dialekt
Seite 458 FM

LINKS: ecView | DTAQ | CabErr

Home | Back | Revert | Apply jurish@bbaw.de

DTA::EvalCorpus::Prune 0.02 / DbCgi version 0.02

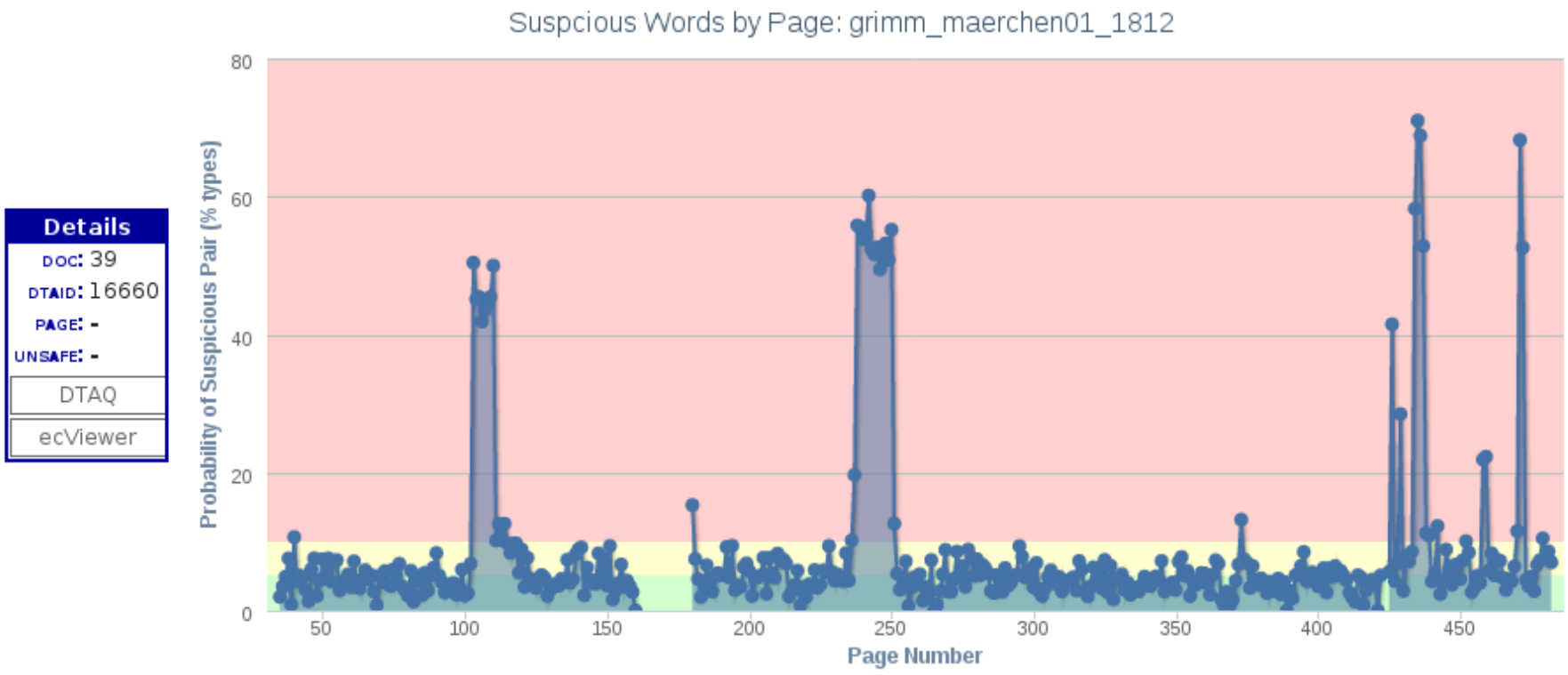
Pruning Tool: Regions

http://kaskade.dwds.de/dta-ecp/edit.perl?doc=39#tabPlot

DTA::EvalCorpus::Prune: Edit (grimm_maerchen01_1812)

Home | Back | | Revert | Apply

Properties | Plot | Pairs



Home | Back | | Revert | Apply

jurish@bbaw.de

DTA::EvalCorpus::Prune 0.02 / DbCgi version 0.02

Pruning Tool: Pairs

http://kaskade.dwds.de/dta-ecp/edit.perl?doc=39#tabPairs

DTA::EvalCorpus::Prune: Edit (grimm_maerchen01_1812)

Home | Back | Revert | Apply

Properties | Plot | Pairs

Suspicious Pairs (1-8 of 2420 total)

First
<< Prev
Next >>

OFFSET: 0
LIMIT: 8

links	rank	wold	wnew	f	status
ecView DTA DTAQ	1	ward	ward	188	-x -X ?m ?M
ecView DTA DTAQ	2	dat	dat	147	-x -X -m -M
ecView DTA DTAQ	3	andern	andern	134	-x -X +m +M
ecView DTA DTAQ	4	ſe	se	103	-x +X ?m ?M
ecView DTA DTAQ	5	een	een	95	-x -X -m -M
ecView DTA DTAQ	6	ſed	sed	92	?x ?X -m -M
ecView DTA DTAQ	7	ſeyn	seyn	83	-x -X -m -M
ecView DTA DTAQ	8	ſey	sey	81	-x -X -m -M

Details for pair 1831

WOLD: andern
WNEW: andern
EXLEX(OLD): anderen (db:486)
EXLEX(NEW): anderen (db:486)
MSAFE(OLD): yes
MSAFE(NEW): yes

CAB(wold)

```
andern
        +[mapclass] -exlex,+id,+xid,+msafe,+moota,-mootxy
        +[xlit] l1=1 lx=1 l1s=andern
        +[lts] \?ande6n <0>
        +[morph] andern[_PIS]\*\*\* <0>
        +[morph/safe] 1
        +[dmoot/tag] andern
        +[dmoot/morph] andern[_PIS]\*\*\* <0>
        +[dmoot/analysis] andern ~ andern <0>
        +[moot/word] andern
        +[moot/tag] ADJA
        +[moot/lemma] andern
        +[moot/analysis] PIS @ andern ~ andern[_PIS]\*\*\* <0>
```

CAB(wnew)

```
andern
        +[mapclass] -exlex,+id,+xid,+msafe,+moota,-mootxy
        +[xlit] l1=1 lx=1 l1s=andern
        +[lts] \?ande6n <0>
        +[morph] andern[_PIS]\*\*\* <0>
        +[morph/safe] 1
        +[dmoot/tag] andern
        +[dmoot/morph] andern[_PIS]\*\*\* <0>
        +[dmoot/analysis] andern ~ andern <0>
        +[moot/word] andern
        +[moot/tag] ADJA
        +[moot/lemma] andern
        +[moot/analysis] PIS @ andern ~ andern[_PIS]\*\*\* <0>
```

Home | Back | Revert | Apply jurish@bbaw.de

DTA::EvalCorpus::Prune 0.02 / DbCgi version 0.02

Corpus Editor: Types View

http://kaskade.dwds.de/dtaec/types.perl?where=wold%3D%27Holle%27

DTA::EvalCorpus::DB: Types

CAB data loaded.

USER: moocow
SELECT: pair,wold,wnew,t0,t1,wclass,gclass, bad,review,seen,pok, pwok, freq, pnotes
FROM: vpair p
WHERE: wold = 'Holle'
GROUP BY:
ORDER BY: freq desc
OFFSET: 0 LIMIT: 10 submit

Home | KWIC (all) | First << Prev Next >> | Revert Apply Legend

Type Record(s) 1-4 of 4

actions	pair	wold	wnew	t0	t1	wclass	gclass	bad	review	seen	pok	pwok	freq	pnotes
DETAILS KWIC	352764	Holle	> Holle		Holle	NAME	a	☐	☐	☐	☑	☑	11	
DETAILS KWIC	289877	Holle	> Hölle		Hölle	LEX	a	☐	☐	☑	☑	☑	1	
DETAILS KWIC	150433	Holle	> Holle		Holle	LEX	a	☐	☐	☐	☑	☐	0	
DETAILS KWIC	~~289876~~	Holle	> Hölle		Hölle	LEX		☑	☐	☐	☐	☐	0	

CAB(OLD)

Holle

EXLEX: Holle (EC/DB:INSERT)
MSAFE: Good
MORPH:
- *Holle* NE FIRSTNAME NONE NONE SG NOM_ACC_DAT <0>
- *Holle* NE LASTNAME NONE K_L_H_M_NAMTI_FAM SG NOM_ACC_DAT <0>
- *Holle* NE GEONAME NONE K_RAUMORT_ART_BODEN_FESTLAND_STADTORT SG NOM_ACC_DAT <0>
- *Holle* NN K_G_DINGNAT_KOERPTEIL FEM SG* <0>

CAB(NEW)

Hölle

EXLEX: Hölle (EC/DB:INSERT)
MSAFE: Good
MORPH:
- *Hölle* NE LASTNAME NONE ABSTR_MYTH_RELIG SG NOM_ACC_DAT <0>
- *Hölle* NN ABSTR FEM SG* <0>
- *Hölle* NN K_G_ARTEF_GEB FEM SG* <0>

Home | KWIC (all) | First << Prev Next >> | Revert Apply jurish@bbaw.de

DTA::EvalCorpus::DB 0.01 / DbCgi version 0.02 0.218462 sec

Corpus Editor: KWIC View

http://kaskade.dwds.de/dtaec/kwic.perl?where=wold%3D%27Holle%27

DTA::EvalCorpus::DB: KWIC

CAB data loaded.

USER: moocow
SELECT: sent,token
FROM: vtoken w
WHERE: wold= 'Holle'
GROUP BY:
ORDER BY:
OFFSET: 0 LIMIT: 5 CONTEXT-WIDTH: 5 submit ☑ old ☑ new

Home | First << Prev Next >> | Revert Apply

Batch Update RESET
+ - --select an attribute-- ▼ =

Legend

Basic
WOLD: Holle
EXLEX(OLD): Holle (EC/DB:INSERT)
MSAFE(OLD): Good
WNEW: Holle
EXLEX(NEW): Holle (EC/DB:INSERT)
MSAFE(NEW): Good

Attributes (read-write)
Attributes (read-only)
Sentence History
Token History

Hit(s) 1-5 of 12

	LEFT	MATCH	RIGHT
DTAQ SENT	... ; ich bin die Frau ... ; ich bin die Frau	Holle Holle	. .
DTAQ SENT	... eine Zeitlang bei der Frau ... eine Zeitlang bei der Frau	Holle Holle	, da ward es traurig ... , da wurde es traurig ...
DTAQ SENT	" Die Frau " Die Frau	Holle Holle	lagte : „ du halt ... sagte : „ du hast ...
DTAQ SENT	... , " sprach die Frau ... , " sprach die Frau	Holle Holle	. .
DTAQ SENT	Als lie vor der Frau Als sie vor der Frau	Holle Holle	Haus kam , fürchtete lie ... Haus kam , fürchtete sie ...

Holle

MORPH:
- Holle NE FIRSTNAME NONE NONE SG NOM_ACC_DAT <0>
- Holle NE LASTNAME NONE K_L_H_M_NAMTI_FAM SG NOM_ACC_DAT <0>
- Holle NE GEONAME NONE K_RAUMORT_ART_BODEN_FESTLAND_STADTORT SG NOM_ACC_DAT <0>
- Holle NN K_G_DINGNAT_KOERPTEIL FEM SG* <0>

Holle

MORPH:
- Holle NE FIRSTNAME NONE NONE SG NOM_ACC_DAT <0>
- Holle NE LASTNAME NONE K_L_H_M_NAMTI_FAM SG NOM_ACC_DAT <0>
- Holle NE GEONAME NONE K_RAUMORT_ART_BODEN_FESTLAND_STADTORT SG NOM_ACC_DAT <0>
- Holle NN K_G_DINGNAT_KOERPTEIL FEM SG* <0>

Home | First << Prev Next >> | Revert Apply jurish@bbaw.de

DTA::EvalCorpus::DB 0.01 / DbCgi v0.02 0.417903 sec

Corpus Editor: Sentence View

http://kaskade.dwds.de/dtaec/sent.perl?sent=86493&token=1823583

DTA::EvalCorpus::DB: Sentence: grimm_maerchen01_1812/s1143 (#86493)

CAB data loaded.

USER: moocow SENT: 86493 CTX: 2 submit

Home | DTAQ | KWIC | << Prev | Next >> | Revert | Apply | Legend

DETAILS

Basic

WOLD:	Holle
EXLEX(OLD):	Holle (EC/DB:INSERT)
MSAFE(OLD):	Good
WNEW:	Holle
EXLEX(NEW):	Holle (EC/DB:INSERT)
MSAFE(NEW):	Good

Attributes (read-write)

WNEW:	Holle
WCLASS:	LEX
POK:	☑
BAD:	☐
REVIEW:	☐
SBAD:	☐
SPLIT:	
SJOIN:	
NOTES:	
PNOTES:	

Attributes (read-only)

Sentence History

Token History

OLD

86491 wir Aepfel ſind allemiteinander reif ! " Da ſchuttelt' es den Baum , daß die Aepfel fielen , als regenten ſie , ſolang bis keiner mehr oben war , darnach ging es wieder fort . 86492 Endlich kam es zu einem kleinen Haus , daraus guckte eine alte Frau , weil ſie aber ſo große Zähne hatte , ward ihm Angſt und es wollte fortlaufen .

86493 Die alte Frau aber rief ihm nach : „ fürcht dich nicht , liebes Kind , bleib bei mir , wenn du alle Arbeit im Haus ordentlich thun willſt , ſo ſoll dirs gut gehn : nur mußt du recht darauf Acht geben daß du

mein Bett gut machſt , und es fleißig aufſchuttelſt , daß die Federn fliegen , dann ſchneit es in der Welt ; ich bin die Frau **Holle** .

86494 Weil die Alte ſo gut ſprach , willigte das Mädchen ein und begab ſich in ihren Dienſt . 86495 Es beſorgte auch alles nach ihrer Zufriedenheit und ſchuttelte ihr das Bett immer gewaltig auf , dafür hatte es auch ein gut Leben bei ihr , kein böſes Wort und alle Tage Geſottenes und Gebratenes .

Holle

MORPH:

Holle	NE	FIRSTNAME NONE NONE SG NOM_ACC_DAT	<0>
Holle	NE	LASTNAME NONE K_L_H_M_NAMTI_FAM SG NOM_ACC_DAT	<0>
Holle	NE	GEONAME NONE K_RAUMORT_ART_BODEN_FESTLAND_STADTORT SG NOM_ACC_DAT	<0>
Holle	NN	K_G_DINGNAT_KOERPTEIL FEM SG*	<0>

NEW

86491 wir Äpfel sind alle miteinander reif ! " Da schüttelte es den Baum , daß die Äpfel fielen , als Regenten sie , solang bis keiner mehr oben war , danach ging es wieder fort . 86492 Endlich kam es zu einem kleinen Haus , daraus guckte eine alte Frau , weil sie aber so große Zähne hatte , wurde ihm Angst und es wollte fortlaufen .

86493 Die alte Frau aber rief ihm nach : „ fürchte dich nicht , liebes Kind , bleibe bei mir , wenn du alle Arbeit im Haus ordentlich tun willst , so soll dir es gut gehn : nur mußt du recht darauf Acht geben daß du

mein Bett gut machst , und es fleißig aufschüttelst , daß die Federn fliegen , dann schneit es in der Welt ; ich bin die Frau **Holle** .

86494 Weil die Alte so gut sprach , willigte das Mädchen ein und begab sich in ihren Dienst . 86495 Es besorgte auch alles nach ihrer Zufriedenheit und schüttelte ihr das Bett immer gewaltig auf , dafür hatte es auch ein gut Leben bei ihr , kein böses Wort und alle Tage Gesottenes und Gebratenes .

Holle

MORPH:

Holle	NE	FIRSTNAME NONE NONE SG NOM_ACC_DAT	<0>
Holle	NE	LASTNAME NONE K_L_H_M_NAMTI_FAM SG NOM_ACC_DAT	<0>
Holle	NE	GEONAME NONE K_RAUMORT_ART_BODEN_FESTLAND_STADTORT SG NOM_ACC_DAT	<0>
Holle	NN	K_G_DINGNAT_KOERPTEIL FEM SG*	<0>

Home | DTAQ | KWIC | << Prev | Next >> | Revert | Apply

jurish@bbaw.de

DTA::EvalCorpus::DB 0.01 / DbCgi v0.02 0.745213 sec