

Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records

Eva Lang[†], Joan Puigcerver[‡], Alejandro H. Toselli[‡] and Enrique Vidal[‡]

[†]Archiv des Bistums Passau
Bischoefliches Ordinariat Passau, Passau, Germany
eva.lang@bistum-passau.de

[‡]Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain
{jpuigcerver,ahector,evidal}@prhlt.upv.es



August 6th, 2018

Outline

- Introduction ▷ 3
- From the Filler Model to Lexicon-free Probabilistic Indexing ▷ 5
- Basic Search and Retrieval (KWS) Results ▷ 6
- Structured Multi-Word Query Search ▷ 8
- Information Extraction from Table Images: Results ▷ 14
- Conclusions ▷ 15

Introduction

- ▶ Huge amounts of legacy handwritten documents exist, but perhaps more than 99.99% of them are *untranscribed*.
- ▶ In particular, text access is in high demand for many *archive* documents: birth, marriage and death records, military draft records, census, property, etc.
Here we deal with a *German handwritten parish record collection* (16th - 19th c.), held by the Passau Diocesan Archives.
- ▶ Rely on **Lexicon-free Probabilistic Indices** (PI) which allow *fast search & retrieval* and other forms of *text data analysis* from untranscribed handwritten text images.
- ▶ Two main contributions of the present work:
 1. Analyze the *impact of transliteration* and *PI density* (size) on indexing and search performance.
 2. Successfully explore the *use of PIs* to support structured, multiple-word queries for *information extraction from untranscribed handwritten tables*.

Lexicon-free Probabilistic Index: Example

0 100 200 300 400 500 600

50. It matters not whether the mis-supposal
 100 regards the matter of fact or matter of law.
 150
 200 The matter of fact where you suppose some.

# pageID="Bentham-071-021-002-part"		REGARDS	0.857	5	115	84	31	THE	0.990	1	198	28	31
# keyword relPrb	bounding box	UGARDS	0.138	5	115	80	31	MATTER	0.934	61	198	64	31
#		THE	0.993	110	115	43	31	OF	0.988	141	198	28	31
2 0.929	1 36 20 31	MATTER	0.998	160	115	93	31	FAST	0.367	182	198	62	31
21 0.064	1 36 24 31	OF	0.996	271	115	23	31	FAR	0.186	182	198	36	31
IT 0.982	33 36 27 31	FACT	0.999	306	115	49	31
IF 0.012	33 36 26 31	OR	0.973	377	115	37	31	FACT	0.017	182	198	46	31
MATTERS 0.989	77 36 99 31	ON	0.021	377	115	42	31	AS	0.142	200	198	29	31
MATTER 0.011	77 36 93 31	MATTER	0.990	425	116	100	31	HAE	0.022	200	198	29	31
NOT 0.999	216 36 7 31	OF	0.995	542	115	25	31	WHERE	0.992	255	198	90	31
WHETHER 1.000	256 36 99 31	LAM	0.407	575	115	30	31	YOU	0.761	365	198	45	31
THE 0.997	389 36 33 31	BIMR	0.175	575	115	55	31	YOW	0.030	365	198	45	31
MIS-SUPPOSAL 1.000	455 36 193 31	GOUS	0.064	372	198	47	31
		LAW	0.032	575	115	36	31	SUPPOSE	0.975	429	198	120	31
THE 0.927	430 88 30 31	TAUE	0.031	575	115	55	31	SUPFROSE	0.024	429	198	125	31
LHE 0.056	434 88 25 31	SOME	0.834	570	198	78	31
...	...	LANE	0.012	575	115	59	31	SONER	0.016	576	198	83	31
								OME	0.109	580	198	65	31
								ME	0.022	620	198	22	31

All character strings or “pseudo-words” which are likely enough to be real words are indexed.

Lexicon-free Probabilistic Index: Example

0 100 200 300 400 500 600

50. 2. It **matter** not whether the **Mis-supposal**
 regards the **matter** of fact or **matter** of law.
 The **matter** of fact where you suppose some-

#	pageID="Bentham-071-021-002-part"	REGARDS	0.857	5	115	84	31	THE	0.990	1	198	28	31
#	keyword relPrb	UGARDS	0.138	5	115	80	31	MATTER	0.934	61	198	64	31
#	bounding box	THE	0.993	110	115	43	31	OF	0.988	141	198	28	31
2	0.929	1	36	20	31			FAST	0.367	182	198	62	31
21	0.064	1	36	24	31			FAR	0.186	182	198	36	31
IT	0.982	33	36	27	31			FACT	0.017	182	198	46	31
IF	0.012	33	36	26	31			AS	0.142	200	198	29	31
MATTERS	0.998	160	115	93	31			HAE	0.022	200	198	29	31
MATTER	0.011	77	36	93	31			WHERE	0.992	255	198	90	31
NOT	0.999	216	36	7	31			YOU	0.761	365	198	45	31
WHETHER	1.000	256	36	99	31			YOW	0.030	365	198	45	31
THE	0.997	389	36	33	31			GOUS	0.064	372	198	47	31
MIS-SUPPOSAL	1.000	455	36	193	31			SUPPOSE	0.975	429	198	120	31
		LAW	0.032	575	115	36	31	SUPFROSE	0.024	429	198	125	31
THE	0.927	430	88	30	31			SOME	0.834	570	198	78	31
LHE	0.056	434	88	25	31			SONER	0.016	576	198	83	31
...	...	TAUE	0.031	575	115	55	31	OME	0.109	580	198	65	31
...	ME	0.022	620	198	22	31

Spots for **MATTER** and **MATTERS** marked in colors according to their Relevance Probabilities.

From the Filler Model to Lexicon-free Probabilistic Indexing

- ▶ Segmentation- & Lexicon-free Filler KWS approaches based on HMM/RNN
 - A. Fischer et al., "Lexicon-free handwritten word spotting using character HMMs" Pattern Recognition Letters, 2012.
 - V. Frinken et al., "A novel word spotting method based on recurrent neural networks" IEEE TPAMI, 2012.
- ▶ Reduce Filler high computing cost using *character lattices* (CL) (same accuracy)
 - A. H. Toselli et al., "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents" ICDAR'13.
- ▶ Filler accuracy improved by adding 2-gram character LM (still much slower)
 - A. Fischer et al., "Improving HMM-Based Keyword Spotting with Character Language Models", ICDAR'13.
- ▶ Use *6-gram LM* to improve Filler accuracy, boost efficiency by means of CLs
 - A. H. Toselli et al., "Context-aware lattice based filler approach for key word spotting in handwritten documents", ICDAR'15.
- ▶ *Filler probabilistic interpretation*: leads to correct spotting *Relevance probability*
 - Puigcerver et al., "Probab. interpret. and improvements to the HMM-filler for handwritten keyword spotting", ICDAR'15.
- ▶ Further improve accuracy and efficiency of probabilistically interpreted Filler model
 - A. H. Toselli et al., "Two methods to improve confidence scores for lexicon-free word spotting in handwritten text" ICFHR'16.
- ▶ *Large-scale Lexicon-free Probabilistic Indexing* (PI) based on the probabilistic Filler
 - T. Bluche et. al., "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project" ICDAR'17.

From the Filler Model to Lexicon-free Probabilistic Indexing: Ours

- ▶ Segmentation- & Lexicon-free Filler KWS approaches based on HMM/RNN
A. Fischer et al., "Lexicon-free handwritten word spotting using character HMMs" Pattern Recognition Letters, 2012.
V. Frinken et al., "A novel word spotting method based on recurrent neural networks" IEEE TPAMI, 2012.
- ▶ Reduce Filler high computing cost using *character lattices* (CL) (same accuracy)
A. H. Toselli et al., "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents" ICDAR'13.
- ▶ Filler accuracy improved by adding 2-gram character LM (still much slower)
A. Fischer et al., "Improving HMM-Based Keyword Spotting with Character Language Models", ICDAR'13.
- ▶ Use *6-gram LM* to improve Filler accuracy, boost efficiency by means of CLs
A. H. Toselli et al., "Context-aware lattice based filler approach for key word spotting in handwritten documents", ICDAR'15.
- ▶ *Filler probabilistic interpretation*: leads to correct spotting *Relevance probability*
Puigcerver et al., "Probab. interpret. and improvements to the HMM-filler for handwritten keyword spotting", ICDAR'15.
- ▶ Further improve accuracy and efficiency of probabilistically interpreted Filler model
A. H. Toselli et al., "Two methods to improve confidence scores for lexicon-free word spotting in handwritten text" ICFHR'16.
- ▶ *Large-scale Lexicon-free Probabilistic Indexing* (PI) based on the probabilistic Filler
T. Bluche et. al., "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project" ICDAR'17.

Lexicon-free Probabilistic Indexing Search Performance: Impact of Transliteration and Language Modeling

Transliteration: normalize spelling, fold diacritics and case of query strings, etc.

Early: when training char. Optical Models; *Late*: when the Probabilistic Index is built

Average Precision (AP), *mean AP* (mAP) for different transliterations and *language models* (LM)

Transliteration	Character LM	AP	mAP
Early	none	0.70	0.66
Early	3-gram	0.71	0.68
Early	6-gram	0.75	0.69
Late	6-gram	0.69	0.66

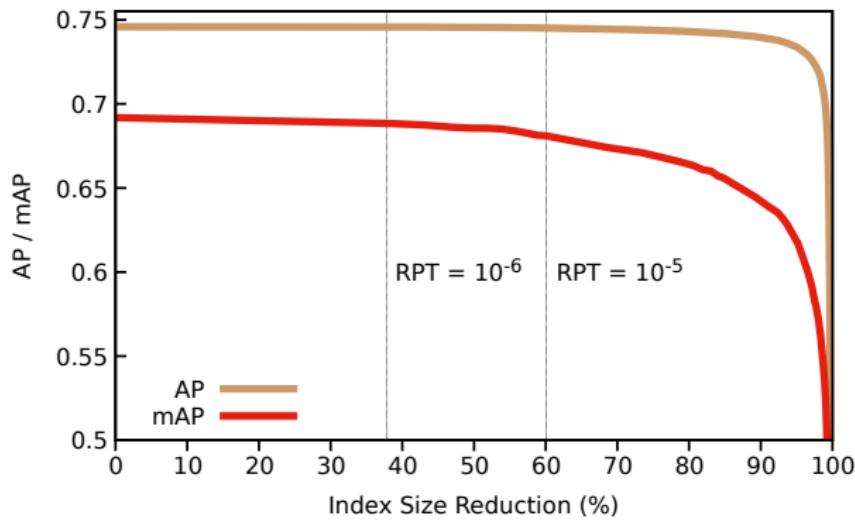
LMs provide useful AP and mAP improvements

Early transliteration proves significantly better

Probabilistic Index Trimming: Effect on Search Performance

Indexing a large number of possibly useless pseudo-word spots do not harm search performance, but do result in large storage overheads, problematic for big collections.

Most unlikely spots can safely be trimmed:



With a Relevance Probability Threshold of 10^{-5} , the average index size per page drops from 56 953 to 22 742 spots (60%), but mAP falls only from 0.69 to 0.68 (and AP decay is negligible)

Outline

- Introduction ▷ 3
- From the Filler Model to Lexicon-free Probabilistic Indexing ▷ 5
- Basic Search and Retrieval (KWS) Results ▷ 6
- Structured Multi-Word Query Search ▷ 8
- Information Extraction from Table Images: Results ▷ 14
- Conclusions ▷ 15

Information Extraction (IE) from Handwritten Table Images

- ▶ Handwritten tables perhaps account for more than half of the vast amounts of documents preserved in archives.
- ▶ Tables contain important, and often ready-to-use information for many historical studies, such as ethnography, demography, economics, genealogy, etc.
- ▶ Accurately transcribing images of handwriting tables is *very difficult*:
 - Ad-hoc, variable, inconsistent and even erratic layouts,
 - difficult line detection,
 - hopeless *reading order* ambiguities,
 - short lines lack linguistic context to help accurate word recognition,
 - etc.

Good news: Probabilistic Indices can support structured, multiple-word queries aiming at complex information extraction from untranscribed images of tabular data.

Towards Information Extraction from Table Images

- ▶ From previous works: *PI's support page-level Boolean multi-word queries*
 - ▶ *PI's hold geometric information* (position, shape and size) of the bounding boxes (BB) of the indexed words
 - ▶ *Boolean queries*, along with BB-based *geometric reasoning*, can be used to support structured queries for *information extraction from table images*

Example of handwritten table images from the *Passau* dataset:

		1919		1920	
Year	Month	Year	Month	Year	Month
1919	Aug	1919	Aug	1920	Aug
1919	Sept	1919	Sept	1920	Sept
1919	Oct	1919	Oct	1920	Oct
1919	Nov	1919	Nov	1920	Nov
1919	Dec	1919	Dec	1920	Dec
1920	Jan	1920	Jan	1921	Jan
1920	Feb	1920	Feb	1921	Feb
1920	Mar	1920	Mar	1921	Mar
1920	Apr	1920	Apr	1921	Apr
1920	May	1920	May	1921	May
1920	June	1920	June	1921	June
1920	July	1920	July	1921	July
1920	Aug	1920	Aug	1921	Aug
1920	Sept	1920	Sept	1921	Sept
1920	Oct	1920	Oct	1921	Oct
1920	Nov	1920	Nov	1921	Nov
1920	Dec	1920	Dec	1921	Dec
1921	Jan	1921	Jan	1922	Jan
1921	Feb	1921	Feb	1922	Feb
1921	Mar	1921	Mar	1922	Mar
1921	Apr	1921	Apr	1922	Apr
1921	May	1921	May	1922	May
1921	June	1921	June	1922	June
1921	July	1921	July	1922	July
1921	Aug	1921	Aug	1922	Aug
1921	Sept	1921	Sept	1922	Sept
1921	Oct	1921	Oct	1922	Oct
1921	Nov	1921	Nov	1922	Nov
1921	Dec	1921	Dec	1922	Dec
1922	Jan	1922	Jan	1923	Jan
1922	Feb	1922	Feb	1923	Feb
1922	Mar	1922	Mar	1923	Mar
1922	Apr	1922	Apr	1923	Apr
1922	May	1922	May	1923	May
1922	June	1922	June	1923	June
1922	July	1922	July	1923	July
1922	Aug	1922	Aug	1923	Aug
1922	Sept	1922	Sept	1923	Sept
1922	Oct	1922	Oct	1923	Oct
1922	Nov	1922	Nov	1923	Nov
1922	Dec	1922	Dec	1923	Dec
1923	Jan	1923	Jan	1924	Jan
1923	Feb	1923	Feb	1924	Feb
1923	Mar	1923	Mar	1924	Mar
1923	Apr	1923	Apr	1924	Apr
1923	May	1923	May	1924	May
1923	June	1923	June	1924	June
1923	July	1923	July	1924	July
1923	Aug	1923	Aug	1924	Aug
1923	Sept	1923	Sept	1924	Sept
1923	Oct	1923	Oct	1924	Oct
1923	Nov	1923	Nov	1924	Nov
1923	Dec	1923	Dec	1924	Dec
1924	Jan	1924	Jan	1925	Jan
1924	Feb	1924	Feb	1925	Feb
1924	Mar	1924	Mar	1925	Mar
1924	Apr	1924	Apr	1925	Apr
1924	May	1924	May	1925	May
1924	June	1924	June	1925	June
1924	July	1924	July	1925	July
1924	Aug	1924	Aug	1925	Aug
1924	Sept	1924	Sept	1925	Sept
1924	Oct	1924	Oct	1925	Oct
1924	Nov	1924	Nov	1925	Nov
1924	Dec	1924	Dec	1925	Dec
1925	Jan	1925	Jan	1926	Jan
1925	Feb	1925	Feb	1926	Feb
1925	Mar	1925	Mar	1926	Mar
1925	Apr	1925	Apr	1926	Apr
1925	May	1925	May	1926	May
1925	June	1925	June	1926	June
1925	July	1925	July	1926	July
1925	Aug	1925	Aug	1926	Aug
1925	Sept	1925	Sept	1926	Sept
1925	Oct	1925	Oct	1926	Oct
1925	Nov	1925	Nov	1926	Nov
1925	Dec	1925	Dec	1926	Dec
1926	Jan	1926	Jan	1927	Jan
1926	Feb	1926	Feb	1927	Feb
1926	Mar	1926	Mar	1927	Mar
1926	Apr	1926	Apr	1927	Apr
1926	May	1926	May	1927	May
1926	June	1926	June	1927	June
1926	July	1926	July	1927	July
1926	Aug	1926	Aug	1927	Aug
1926	Sept	1926	Sept	1927	Sept
1926	Oct	1926	Oct	1927	Oct
1926	Nov	1926	Nov	1927	Nov
1926	Dec	1926	Dec	1927	Dec
1927	Jan	1927	Jan	1928	Jan
1927	Feb	1927	Feb	1928	Feb
1927	Mar	1927	Mar	1928	Mar
1927	Apr	1927	Apr	1928	Apr
1927	May	1927	May	1928	May
1927	June	1927	June	1928	June
1927	July	1927	July	1928	July
1927	Aug	1927	Aug	1928	Aug
1927	Sept	1927	Sept	1928	Sept
1927	Oct	1927	Oct	1928	Oct
1927	Nov	1927	Nov	1928	Nov
1927	Dec	1927	Dec	1928	Dec
1928	Jan	1928	Jan	1929	Jan
1928	Feb	1928	Feb	1929	Feb
1928	Mar	1928	Mar	1929	Mar
1928	Apr	1928	Apr	1929	Apr
1928	May	1928	May	1929	May
1928	June	1928	June	1929	June
1928	July	1928	July	1929	July
1928	Aug	1928	Aug	1929	Aug
1928	Sept	1928	Sept	1929	Sept
1928	Oct	1928	Oct	1929	Oct
1928	Nov	1928	Nov	1929	Nov
1928	Dec	1928	Dec	1929	Dec
1929	Jan	1929	Jan	1930	Jan
1929	Feb	1929	Feb	1930	Feb
1929	Mar	1929	Mar	1930	Mar
1929	Apr	1929	Apr	1930	Apr
1929	May	1929	May	1930	May
1929	June	1929	June	1930	June
1929	July	1929	July	1930	July
1929	Aug	1929	Aug	1930	Aug
1929	Sept	1929	Sept	1930	Sept
1929	Oct	1929	Oct	1930	Oct
1929	Nov	1929	Nov	1930	Nov
1929	Dec	1929	Dec	1930	Dec
1930	Jan	1930	Jan	1931	Jan
1930	Feb	1930	Feb	1931	Feb
1930	Mar	1930	Mar	1931	Mar
1930	Apr	1930	Apr	1931	Apr
1930	May	1930	May	1931	May
1930	June	1930	June	1931	June
1930	July	1930	July	1931	July
1930	Aug	1930	Aug	1931	Aug
1930	Sept	1930	Sept	1931	Sept
1930	Oct	1930	Oct	1931	Oct
1930	Nov	1930	Nov	1931	Nov
1930	Dec	1930	Dec	1931	Dec
1931	Jan	1931	Jan	1932	Jan
1931	Feb	1931	Feb	1932	Feb
1931	Mar	1931	Mar	1932	Mar
1931	Apr	1931	Apr	1932	Apr
1931	May	1931	May	1932	May
1931	June	1931	June	1932	June
1931	July	1931	July	1932	July
1931	Aug	1931	Aug	1932	Aug
1931	Sept	1931	Sept	1932	Sept
1931	Oct	1931	Oct	1932	Oct
1931	Nov	1931	Nov	1932	Nov
1931	Dec	1931	Dec	1932	Dec
1932	Jan	1932	Jan	1933	Jan
1932	Feb	1932	Feb	1933	Feb
1932	Mar	1932	Mar	1933	Mar
1932	Apr	1932	Apr	1933	Apr
1932	May	1932	May	1933	May
1932	June	1932	June	1933	June
1932	July	1932	July	1933	July
1932	Aug	1932	Aug	1933	Aug
1932	Sept	1932	Sept	1933	Sept
1932	Oct	1932	Oct	1933	Oct
1932	Nov	1932	Nov	1933	Nov
1932	Dec	1932	Dec	1933	Dec
1933	Jan	1933	Jan	1934	Jan
1933	Feb	1933	Feb	1934	Feb
1933	Mar	1933	Mar	1934	Mar
1933	Apr	1933	Apr	1934	Apr
1933	May	1933	May	1934	May
1933	June	1933	June	1934	June
1933	July	1933	July	1934	July
1933	Aug	1933	Aug	1934	Aug
1933	Sept	1933	Sept	1934	Sept
1933	Oct	1933	Oct	1934	Oct
1933	Nov	1933	Nov	1934	Nov
1933	Dec	1933	Dec	1934	Dec
1934	Jan	1934	Jan	1935	Jan
1934	Feb	1934	Feb	1935	Feb
1934	Mar	1934	Mar	1935	Mar
1934	Apr	1934	Apr	1935	Apr
1934	May	1934	May	1935	May
1934	June	1934	June	1935	June
1934	July	1934	July	1935	July
1934	Aug	1934	Aug	1935	Aug
1934	Sept	1934	Sept	1935	Sept
1934	Oct	1934	Oct	1935	Oct
1934	Nov	1934	Nov	1935	Nov
1934	Dec	1934	Dec	1935	Dec
1935	Jan	1935	Jan	1936	Jan
1935	Feb	1935	Feb	1936	Feb
1935	Mar	1935	Mar	1936	Mar
1935	Apr	1935	Apr	1936	Apr
1935	May	1935	May	1936	May
1935	June	1935	June	1936	June
1935	July	1935	July	1936	July
1935	Aug	1935	Aug	1936	Aug
1935	Sept	1935	Sept	1936	Sept
1935	Oct	1935	Oct	1936	Oct
1935	Nov	1935	Nov	1936	Nov
1935	Dec	1935	Dec	1936	Dec
1936	Jan	1936	Jan	1937	Jan
1936	Feb	1936	Feb	1937	Feb
1936	Mar	1936	Mar	1937	Mar
1936	Apr	1936	Apr	1937	Apr
1936	May	1936	May	1937	May
1936	June	1936	June	1937	June
1936	July	1936	July	1937	July
1936	Aug	1936	Aug	1937	Aug
1936	Sept	1936	Sept	1937	Sept
1936	Oct	1936	Oct	1937	Oct
1936	Nov	1936	Nov	1937	Nov
1936	Dec	1936	Dec	1937	Dec
1937	Jan	1937	Jan	1938	Jan
1937	Feb	1937	Feb	1938	Feb
1937	Mar	1937	Mar	1938	Mar
1937	Apr	1937	Apr	1938	Apr
1937	May	1937	May	1938	May
1937	June	1937	June	1938	June
1937	July	1937	July	1938	July
1937	Aug	1937	Aug	1938	Aug
1937	Sept	1937	Sept	1938	Sept
1937	Oct	1937	Oct	1938	Oct
1937	Nov	1937	Nov	1938	Nov
1937	Dec	1937	Dec	1938	Dec
1938	Jan	1938	Jan	1939	Jan
1938	Feb	1938	Feb	1939	Feb
1938	Mar	1938	Mar	1939	Mar
1938	Apr	1938	Apr	1939	Apr
1938	May	1938	May	1939	May
1938	June	1938	June	1939	June
1938	July	1938	July	1939	July
1938	Aug	1938	Aug	1939	Aug
1938	Sept	1938	Sept	1939	Sept
1938	Oct	1938	Oct	1939	Oct
1938	Nov	1938	Nov	1939	Nov
1938	Dec	1938	Dec	1939	Dec
1939	Jan	1939	Jan	1940	Jan
1939	Feb	1939	Feb	1940	Feb
1939	Mar	1939	Mar	1940	Mar
1939	Apr	1939	Apr	1940	Apr
1939	May	1939	May	1940	May
1939	June	1939	June	1940	June
1939	July	1939	July	1940	July
1939	Aug	1939	Aug	1940	Aug
1939	Sept	1939	Sept	1940	Sept
1939	Oct	1939	Oct	1940	Oct
1939	Nov	1939	Nov	1940	Nov
1939					

Structured Multi-word Queries for IE from Table Images

Aim to deal with queries of the form:

`(column-heading, column-content)`

where **column-heading** is an AND combination of table heading words and **column-content** is a (single) keyword.[†]

Examples:

- ▶ `(ORT, STEINERLEINBACH)` ((PLACE, STEINERLEINBACH))
- ▶ `(TAUF TAG, APRIL)` ((BAPTISM DAY, APRIL))
- ▶ `(KRANKHEIT ARZT, FRAISEN)` ((CAUSE OF DEATH, SPASMS))
- ▶ `(NAMEN DES BRAEUTIGAMS, JOSEF)` ((NAME OF THE GROOM, JOSEF))
- ▶ `(NAMEN DER BRAUT, MARIA)` ((NAME OF THE BRIDE, MARIA))
- ▶ `(TAG MONAT JAHR TODES, 1879)` ((DAY MONTH YEAR OF DEATH, 1879))

[†] More complex structured queries can be similarly supported

Probabilistic Framework for Structured Multi-word Query Search

- ▶ Let $\mathbf{h} \stackrel{\text{def}}{=} \{h_1, h_2 \dots, h_I\}$ be the set of column-heading query words and let $R(h_i)$ denote the *relevance probability* (RP) of h_i ($h_i \in \{0, 1\}$ relevant Boolean variable).
- ▶ Let s_{i1}, \dots, s_{iJ_i} denote the $J_i \geq 1$ different spots of h_i and $R(s_{ij}) \stackrel{\text{def}}{=} P(R | \mathbf{c}_{h_i}, \mathbf{x}_{ij})$ its RP in the image location \mathbf{x}_{ij} , where \mathbf{c}_{h_i} is the character spelling of the word h_i .
- ▶ Then, the RP of the AND combination for the words in \mathbf{h} is computed as:

$$R(\mathbf{h}) = R(h_1 \wedge h_2 \wedge \dots \wedge h_I) \approx \min_{1 \leq i \leq I} R(h_i) \approx \min_{1 \leq i \leq I} \max_{1 \leq j \leq J_i} R(s_{ij}) \quad (\text{see } \dagger)$$

- ▶ Let v_1, \dots, v_K , $K \geq 1$, be the different spots of v retrieved in column locations $\mathbf{x}_1, \dots, \mathbf{x}_K$ and let $R(v_k) \stackrel{\text{def}}{=} P(R | \mathbf{c}_v, \mathbf{x}_k)$ be the RP of the k -th spot.
- ▶ The RP of the column-content word v in the considered column is computed as:

$$R(v) \approx \max_{1 \leq k \leq K} R(v_k)$$

- ▶ Finally, the RP of a column-wise structured multi-word query is computed as:

$$R(\langle \mathbf{h}, v \rangle) = R(\mathbf{h} \wedge v) \approx \min(R(\mathbf{h}), R(v)) \quad (\text{see } \dagger)$$

[†] A.H. Toselli, E. Vidal, J. Puigcerver and E. Noya-García: Probabilistic Multi-Word Spotting in Handwritten Text Images. To be published in Journal of Pattern Analysis and Applications. 2018.

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

User query

Lanigroßf. bisfrigeren. Aufsonstalts- Cst.	Eltern mit Bauß. in die für. namen. sind bei der Mutter der Sohn auf der Gräberplatte; Name.	Lebig. wie Mitt. wurden beobachtet dafür. Es ist No. meine Schwester. beneige Mamas. bei einer Gr. Abendessen. Es Name. Es' auswählen Mamas?	Grbecken nach mit einer Gr. Abendessen Name?	verschafft. vor dem in den Ball. unternehmen.	Zerig. mit Bauß. in die für. namen; Name im Aufzug fallen.	Gräberplatte; zwei von Vierjahr. in der Gedächtnis mit einer offne Dominikanerinnen.	Getragen mit einer offne Vierjahr. in der Gedächtnis mit einer offne Dominikanerinnen.
Aufsonstalts- Cst.	Judith Günther H. Maria Ansgar +	Lebig 1837 2. Februar 1837 Zwischen 24.3. 1836	Katharina. Günther Pfarrer Katharina Lilienberg Lilienberg	Joh. Günther Lilienberg Joh. Günther Lilienberg Lilienberg	2. März Maria	Am 11. Januar 1837	Am 11. Januar 1837
	Maria	1822		Maria. Hochzeit			

Example

Query: < NAMEN DER BRAUT, MARIA >

Spotting heading words

Lamigroßfl. bisfornijen Anfangsfall. Cet.	Blume mit Raufärb' Jü- nison. inn bei der Mittwoch der Sonntags auf der Großeckfl. Name.	Lang. von Mitt. vor im letzten Jahr ^{die} der Blü- me zu wachsen, braun Mammi; bei einer Rei- he fürstlichem Da- knamo. Bei zweyem Mammi.	Geblieben nur nur nur nur?	Der Har. nur nur Leyton Hall. nur nur?	Zweigen mit Raufärb' Jü- nison. Name in einer Rei- se färbt.	Graswuchs. zwei, wie feilt nur nur inn nur?	Geblau mit einer grüne Wippe; in der Grasen mit einer grüne dominativen.
Wolfsbl.							
Blauegloßfl. Lippenbl.	Jahrel Blauegloßfl. -3. J. 1834 H. Mammo zub. Augenwurz +	Lady	1837 1837 1837 Zwischen 24.3. 1836	Gebüsch. Blauegloßfl. Pflanze	Wolfsbl. Blauegloßfl. Lippenbl. Blauegloßfl.	Wolfsbl. Hain	Wolfsbl. Hain
	Mutteria		1822		Mutter. Pflanzung		17

$h_1 = \text{NAMEN}$

$$h_2 = \text{DER}$$

$h_3 = \text{BRAUT}$

Example

Query: < NAMEN DER BRAUT, MARIA >

Spotting heading words

Lamigroßfisch, bis zu 1 m. Leben auf der Meeresfläche.	Eltern mit Küpfchen & Ju- nior. Jun- ior von Mutter abgespalten auf die Meeresfläche. Name.	Lang vor Welt- vor im letzten Stile zu Was- mer im anderen braun Männer; bei einer Ge- schwisterin da Name. Bei zweijährigen Männern.	Gebooren vor dem ab den Ball unter Wasser.	vor Was- mer von dem Ball unter Wasser.	Zwischen mit Küpfchen & Ju- nior von dem Ball unter Wasser. Name & Anzahl bekannt.	Gründel mit Küpfchen & Ju- nior von dem Ball unter Wasser. Name & Anzahl bekannt.	Getrennt mit einer offene Wippe; in den Gründen mit einer offne Verbindungswand.
Gründel Grauflanke abgetrennt von der Mutter zur Augenwelle +	Gründel Grauflanke abgetrennt von der Mutter zur Augenwelle +	1837 1837 1837 1836	Gebooren. abtrennen Pflanze	1837 abtrennen Pflanze	1837 abtrennen Pflanze	1837 abtrennen Pflanze	1837 abtrennen Pflanze

$h_1 = \text{NAMEN}$

$$h_2 = \text{DER}$$

$h_3 = \text{BRAUT}$

Example

Query: { NAMEN DER BRAUT, MARIA }

Applying geometric restrictions

Lanigroßl., bischofsgesell. Aufschiffung Ort	Flurme mit Rauf. in e. für. namen. inne bei den Mittwoch im Land auf im Gräflichstift; Name.	Lanig. der Will. vor ein Dach Süd. der Na. me des zweiten. bonum Mannus. bei einer Ge. schriftsteller. So. Name. Sei ausgewiesen Mannus.	Grbeoren vor vero mitten	vorffas. vor vero in den Rott. vor verber.	Zonigen mit Rauf. in e. für. namen. Name im Aufzuge falliert.	Grbeoren. zwei; vor Zonigen; in der Judent mit einer offna dominikanischen.	Grbeoren mit einer offna Zonigen; in der Judent mit einer offna dominikanischen.
Aufschiffung Lippe	Flurme Gruning. H. a. man in Angebot +	1837 2. 7. 30. 1837 Zwischen 2. 6. 1836	Flurme. Gruning. Rheine	Flurme. Gruning. Lippe Lippe Lippe	Flurme. Gruning. Lippe Lippe Lippe	Flurme. Gruning. Lippe Lippe Lippe	Flurme. Gruning. Lippe
Maria		1822	Maria. Hochzeit				

$h_1 = \text{NAMEN}$

$h_2 = \text{DER}$

$h_3 = \text{BRAUT}$

Relevance Prob.: $R(\mathbf{h}) \approx \min(R(h_1), R(h_2), R(h_3))$

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩ Candidate regions for column-content words

Lanigroßf. bisfranzos. Aufschliff. Ort.	Blumen mit Bauern für namen, eine bei der Mutter der Landt auf der Gräfinfeste Name.	Lebzig wird Mitt. nur im letzten Stil der Frau meine Schwester, bessere Mamas. bei einer Gr. für einigen der Name. Bei einer Mamas?	Grbeeren	vor offen vor vere vor den Ball. vor und vor unterbroch.	Zwischen mit Bauern für namen; Name im Aufzuge falsiert.	Grünaffels; zwei von Vorlagen; im der Farben mit einer offne dominativen.	Getrocknet mit oder ohne Vorlagen; im der Farben mit einer offne dominativen.
Aufschliff	Julius Günther H. Müller Augustus +	1837 2. Februar 1837 Zwischen 24.3. 1836	Katharina Julius Katharina Rheine	Zwischen Ludwig Johann Ludwig Ludwig	Julius Günther Ludwig Johann Ludwig	Julius Günther Ludwig Johann Ludwig	Julius Katharina
Maria		1822		Maria Theresia			17

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

Spotting column-content words

Lamigroßt. bislang aufgewallt; Cob.	Blume mit Raufärbte zu- nehmen; im bei der Mittler der Sonne auf der Grasfläche; Name:	Sehr großer Blatt, vor, in leichter Stille. Bei Stu- mme bei weicher blumen Männer; bei einer Ge- schwindigkeit der Name: Bei weichem Mannen.	Geblüht	Vorffas. vor dem raum	Zwischen mit Raufärbte zu- nehmen; Blume mit Anfang sehr schwach. fallen.	Granaffeli; zweit, an grill von dem Graden mit einer gro- ßen Konzentration.
Wulffbau	Jahrel Ruhm Lippen;	1837	Kathar.	Jahrl. Ruhm Lippen	Wulff. Kathar.	Wulff. Kathar.
	1837 Jahrel Ruhm Lippen;	1837	Dahlinge	1837 Jahrel Ruhm Lippen	Wulff. Kathar.	Wulff. Kathar.
	1837 Jahrel Ruhm Lippen;	1836	Phaner	1836 Jahrel Ruhm Lippen	Wulff. Kathar.	Wulff. Kathar.

$v = \text{MARIA}$ Relevance Prob.: $R(v) \approx \max(R(v_1), R(v_2))$

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

Retrieved spot and its relevance probability

Lanzegeöfft, bisfroniger. Aufschlüssel. Cet.	Ullens mit Rauf-mie für namen, in bei den Mutter der Spur auf der Geistigheit; Name.	Leidig war Will vor ein Dörfchen dalle von Ma me im Lande. boner Mannen. bei einer Ge stigheit war Name. Sp' erweigen Mannen?	Grbeoren vor vere ersten Roll unterbrac	vor öffne Zonigen mit Rauf-mie für namen, Name im Aufste faliert.	Grinna filii zand, vor Vigjons, in den Gelehr mit einer offna dominatioen.
Alles best Almung f Lippen	Ja haw Gruning ing H. a. manne geb. Angrunne +	Leidig 1837 2. J. 26. 1837 Zwischen 24.3. 1836	Kathar. Leidig 1837 Kathar. Leidig Zwischen 24.3. 1836	Joh. Grünig Lippen. Joh. Leidig Leidig Lippen Lippen f. Leidig Kathar.	Ja haw Kathar. Leidig 2. J. 26. Kathar. Leidig Leidig Kathar.

$h_1 = \text{NAMEN}$

$h_2 = \text{DER}$

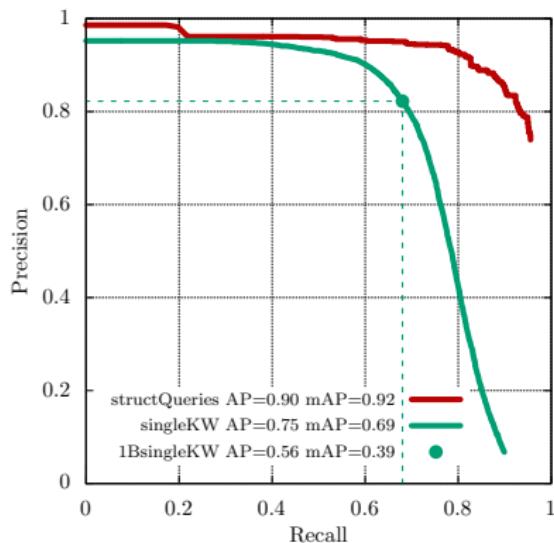
$h_3 = \text{BRAUT}$

Relevance Prob.: $R(\langle \mathbf{h}, v \rangle) = R(\mathbf{h} \wedge v) \approx \min(R(\mathbf{h}), R(v))$

$v = \text{MARIA}$

Information Extraction from Handwritten Table Images: Results

Search performance for single and structured word queries:



Dataset training and test details

- ▶ **PASSAU:** German/Latin, *many hands*.
Training: 200 pages, 102 char CRNN OMs
+ char 6-gram LM trained on training transcripts;
Lexicon: 12 381 tokens.
Test: 91 page images; *Query set:* 6 500 keywords
- ▶ **PASSAUSTRUC:** Table queries in PASSAU.
Training: same as PASSAU. *Test:* 44 table images;
Query set: 363 real multi-word structured queries.
- ▶ See:
<http://transcriptorium.eu/demots/kws-Passau>

Outstanding table information extraction results based on multi-word structured queries

Conclusions

- ▶ The present work confirms with another difficult collection the high effectiveness of lexicon-free single-keyword KWS supported by Probabilistic Indices
- ▶ PIs have been shown to support structured queries involving many words, which allow for complex information retrieval in text images containing tabular data
- ▶ Empirical results validate the proposed approaches for actually indexing the full *Passau* collection, with more than 800 000 historical handwritten register images
- ▶ A real demonstrator of the indexing and search techniques developed and evaluated in this work is publicly available at:

<http://transcriptorium.eu/demots/kws-Passau>

(no yet supporting table information extraction queries)

Thanks for Your Attention !

PASSAU Collection and Experimental Dataset

XVI-XVIII century collection of historical records. 26,000 images, written in German.

		Jahr	
1	Person	Birth	Death
2	Person	Baptism	Marriage
3	Person	Death	Burial
4	Person	Baptism	Burial
5	Person	Baptism	Burial
6	Person	Baptism	Burial
7	Person	Baptism	Burial
8	Person	Baptism	Burial
9	Person	Baptism	Burial
10	Person	Baptism	Burial
11	Person	Baptism	Burial
12	Person	Baptism	Burial
13	Person	Baptism	Burial
14	Person	Baptism	Burial
15	Person	Baptism	Burial
16	Person	Baptism	Burial
17	Person	Baptism	Burial
18	Person	Baptism	Burial
19	Person	Baptism	Burial
20	Person	Baptism	Burial
21	Person	Baptism	Burial
22	Person	Baptism	Burial
23	Person	Baptism	Burial
24	Person	Baptism	Burial
25	Person	Baptism	Burial
26	Person	Baptism	Burial
27	Person	Baptism	Burial
28	Person	Baptism	Burial
29	Person	Baptism	Burial
30	Person	Baptism	Burial
31	Person	Baptism	Burial
32	Person	Baptism	Burial
33	Person	Baptism	Burial
34	Person	Baptism	Burial
35	Person	Baptism	Burial
36	Person	Baptism	Burial
37	Person	Baptism	Burial
38	Person	Baptism	Burial
39	Person	Baptism	Burial
40	Person	Baptism	Burial
41	Person	Baptism	Burial
42	Person	Baptism	Burial
43	Person	Baptism	Burial
44	Person	Baptism	Burial
45	Person	Baptism	Burial
46	Person	Baptism	Burial
47	Person	Baptism	Burial
48	Person	Baptism	Burial
49	Person	Baptism	Burial
50	Person	Baptism	Burial
51	Person	Baptism	Burial
52	Person	Baptism	Burial
53	Person	Baptism	Burial
54	Person	Baptism	Burial
55	Person	Baptism	Burial
56	Person	Baptism	Burial
57	Person	Baptism	Burial
58	Person	Baptism	Burial
59	Person	Baptism	Burial
60	Person	Baptism	Burial
61	Person	Baptism	Burial
62	Person	Baptism	Burial
63	Person	Baptism	Burial
64	Person	Baptism	Burial
65	Person	Baptism	Burial
66	Person	Baptism	Burial
67	Person	Baptism	Burial
68	Person	Baptism	Burial
69	Person	Baptism	Burial
70	Person	Baptism	Burial
71	Person	Baptism	Burial
72	Person	Baptism	Burial
73	Person	Baptism	Burial
74	Person	Baptism	Burial
75	Person	Baptism	Burial
76	Person	Baptism	Burial
77	Person	Baptism	Burial
78	Person	Baptism	Burial
79	Person	Baptism	Burial
80	Person	Baptism	Burial
81	Person	Baptism	Burial
82	Person	Baptism	Burial
83	Person	Baptism	Burial
84	Person	Baptism	Burial
85	Person	Baptism	Burial
86	Person	Baptism	Burial
87	Person	Baptism	Burial
88	Person	Baptism	Burial
89	Person	Baptism	Burial
90	Person	Baptism	Burial
91	Person	Baptism	Burial
92	Person	Baptism	Burial
93	Person	Baptism	Burial
94	Person	Baptism	Burial
95	Person	Baptism	Burial
96	Person	Baptism	Burial
97	Person	Baptism	Burial
98	Person	Baptism	Burial
99	Person	Baptism	Burial
100	Person	Baptism	Burial
101	Person	Baptism	Burial
102	Person	Baptism	Burial
103	Person	Baptism	Burial
104	Person	Baptism	Burial
105	Person	Baptism	Burial
106	Person	Baptism	Burial
107	Person	Baptism	Burial
108	Person	Baptism	Burial
109	Person	Baptism	Burial
110	Person	Baptism	Burial
111	Person	Baptism	Burial
112	Person	Baptism	Burial
113	Person	Baptism	Burial
114	Person	Baptism	Burial
115	Person	Baptism	Burial
116	Person	Baptism	Burial
117	Person	Baptism	Burial
118	Person	Baptism	Burial
119	Person	Baptism	Burial
120	Person	Baptism	Burial
121	Person	Baptism	Burial
122	Person	Baptism	Burial
123	Person	Baptism	Burial
124	Person	Baptism	Burial
125	Person	Baptism	Burial
126	Person	Baptism	Burial
127	Person	Baptism	Burial
128	Person	Baptism	Burial
129	Person	Baptism	Burial
130	Person	Baptism	Burial
131	Person	Baptism	Burial
132	Person	Baptism	Burial
133	Person	Baptism	Burial
134	Person	Baptism	Burial
135	Person	Baptism	Burial
136	Person	Baptism	Burial
137	Person	Baptism	Burial
138	Person	Baptism	Burial
139	Person	Baptism	Burial
140	Person	Baptism	Burial
141	Person	Baptism	Burial
142	Person	Baptism	Burial
143	Person	Baptism	Burial
144	Person	Baptism	Burial
145	Person	Baptism	Burial
146	Person	Baptism	Burial
147	Person	Baptism	Burial
148	Person	Baptism	Burial
149	Person	Baptism	Burial
150	Person	Baptism	Burial
151	Person	Baptism	Burial
152	Person	Baptism	Burial
153	Person	Baptism	Burial
154	Person	Baptism	Burial
155	Person	Baptism	Burial
156	Person	Baptism	Burial
157	Person	Baptism	Burial
158	Person	Baptism	Burial
159	Person	Baptism	Burial
160	Person	Baptism	Burial
161	Person	Baptism	Burial
162	Person	Baptism	Burial
163	Person	Baptism	Burial
164	Person	Baptism	Burial
165	Person	Baptism	Burial
166	Person	Baptism	Burial
167	Person	Baptism	Burial
168	Person	Baptism	Burial
169	Person	Baptism	Burial
170	Person	Baptism	Burial
171	Person	Baptism	Burial
172	Person	Baptism	Burial
173	Person	Baptism	Burial
174	Person	Baptism	Burial
175	Person	Baptism	Burial
176	Person	Baptism	Burial
177	Person	Baptism	Burial
178	Person	Baptism	Burial
179	Person	Baptism	Burial
180	Person	Baptism	Burial
181	Person	Baptism	Burial
182	Person	Baptism	Burial
183	Person	Baptism	Burial
184	Person	Baptism	Burial
185	Person	Baptism	Burial
186	Person	Baptism	Burial
187	Person	Baptism	Burial
188	Person	Baptism	Burial
189	Person	Baptism	Burial
190	Person	Baptism	Burial
191	Person	Baptism	Burial
192	Person	Baptism	Burial
193	Person	Baptism	Burial
194	Person	Baptism	Burial
195	Person	Baptism	Burial
196	Person	Baptism	Burial
197	Person	Baptism	Burial
198	Person	Baptism	Burial
199	Person	Baptism	Burial
200	Person	Baptism	Burial
201	Person	Baptism	Burial
202	Person	Baptism	Burial
203	Person	Baptism	Burial
204	Person	Baptism	Burial
205	Person	Baptism	Burial
206	Person	Baptism	Burial
207	Person	Baptism	Burial
208	Person	Baptism	Burial
209	Person	Baptism	Burial
210	Person	Baptism	Burial
211	Person	Baptism	Burial
212	Person	Baptism	Burial
213	Person	Baptism	Burial
214	Person	Baptism	Burial
215	Person	Baptism	Burial
216	Person	Baptism	Burial
217	Person	Baptism	Burial
218	Person	Baptism	Burial
219	Person	Baptism	Burial
220	Person	Baptism	Burial
221	Person	Baptism	Burial
222	Person	Baptism	Burial
223	Person	Baptism	Burial
224	Person	Baptism	Burial
225	Person	Baptism	Burial
226	Person	Baptism	Burial
227	Person	Baptism	Burial
228	Person	Baptism	Burial
229	Person	Baptism	Burial
230	Person	Baptism	Burial
231	Person	Baptism	Burial
232	Person	Baptism	Burial
233	Person	Baptism	Burial
234	Person	Baptism	Burial
235	Person	Baptism	Burial
236	Person	Baptism	Burial
237	Person	Baptism	Burial
238	Person	Baptism	Burial
239	Person	Baptism	Burial
240	Person	Baptism	Burial
241	Person	Baptism	Burial
242	Person	Baptism	Burial
243	Person	Baptism	Burial
244	Person	Baptism	Burial
245	Person	Baptism	Burial
246	Person	Baptism	Burial
247	Person	Baptism	Burial
248	Person	Baptism	Burial
249	Person	Baptism	Burial
250	Person	Baptism	Burial
251	Person	Baptism	Burial
252	Person	Baptism	Burial
253	Person	Baptism	Burial
254	Person	Baptism	Burial
255	Person	Baptism	Burial
256	Person	Baptism	Burial
257	Person	Baptism	Burial
258	Person	Baptism	Burial
259	Person	Baptism	Burial
260	Person	Baptism	Burial
261	Person	Baptism	Burial
262	Person	Baptism	Burial
263	Person	Baptism	Burial
264	Person	Baptism	Burial
265	Person	Baptism	Burial
266	Person	Baptism	Burial
267	Person	Baptism	Burial
268	Person	Baptism	Burial
269	Person	Baptism	Burial
270	Person	Baptism	Burial
271	Person	Baptism	Burial
272	Person	Baptism	Burial
273	Person	Baptism	Burial
274	Person	Baptism	Burial
275	Person	Baptism	Burial
276	Person	Baptism	Burial
277	Person	Baptism	Burial
278	Person	Baptism	Burial
279	Person	Baptism	Burial
280	Person	Baptism	Burial
281	Person	Baptism	Burial
282	Person	Baptism	Burial
283	Person	Baptism	Burial
284	Person	Baptism	Burial
285	Person	Baptism	Burial
286	Person	Baptism	Burial
287	Person	Baptism	Burial
288	Person	Baptism	Burial
289	Person	Baptism	Burial
290	Person	Baptism	Burial
291	Person	Baptism	Burial



Information about the baptized, married and die parishioners of the various Passau's Diocese parishes.

	Train+Val	Test	TabTest
#Pages	200	89	44
#Lines	29 314	16 376	11 710
RWs	72 848	37 354	21 027
RWs w/o PMs	—	26 709	15 141
Lex-size	12 381	6 532	3 455
#Chars	220	187	119
Transl. Lex-size*	11 160	5 801	3 141
#Transl. Chars*	99	87	73

Statistics for single and structured word queries.
Complex, varied layout, many tables, etc.:
2.4 average words/line ⇒ low LM impact.

* *Transliteration*: all chars uppercase, no diacritics, non-ASCII symbols mapped to ASCII "equivalents"

Probabilistic Index Size and Transliteration

Probabilistic Indices:

- ▶ may become huge (large amounts of storage) for vast manuscript collections.
- ▶ contain large quantities of pseudo-words which probably will never be spotted.

Solution: PI size reduction through filtering out entries whose relevance probability scores fall below a specified threshold.

The medieval German record collection used here contains:

- ▶ different spelling variations of the same word (e.g., accents, umlauts, tie bar, ...)
- ▶ 263 UTF-8 different symbols, most of which are/contain non-ASCII characters.
- ▶ most of such characters can not be typed on standard keyboards.

Solution: Every char/symbol is transliterated by case folding and by removing diacritics and mapping non-ASCII symbols onto their ASCII equivalents.

Remov.	Diacrit.		Non-ASCII to ASCII					
č, č, č	C		Æ, æ	AE	ij	II	ŋ	EN
è, ê, ë	E		Œ, œ	OE	ß	SS	g	US
ṁ, ţ, ţ	M		p, ſ	PRO	d	DE	ð	DER

The benefits are two-fold: a) simplify the composition of queries and b) avoid the waste of probability mass which often leads to degrade search performance.

PI performance: Impact of Transliteration and Language Modeling

Transliteration: normalize spelling and fold diacritics and case of query strings.

Early: at the of optical modelling; *Late*: after the PI is built

Transliteration	Latt-type	Char LM	AP	mAP	MxRc ₁₀
Early	CLs	none	0.701	0.661	0.861
Early	CLs	3-gram	0.712	0.677	0.876
Early	CLs	6-gram	0.746	0.692	0.886
Late	CLs	6-gram	0.692	0.662	0.854
Early	1-best	6-gram	0.559	0.387	0.680
Late	1-best	6-gram	0.492	0.331	0.613

Average Precision (AP), mean AP (mAP) and maximum recall at 10% precision (MxRc₁₀) for different character lattices and language models (LM).

Average Precision (AP) versus Mean Average Precision (mAP)

	AP	mAP
Rank type	Global	Local
Averaging type	Micro: over all query events	Macro: over the APs of isolated queries
Score consistency impact	yes	no
Demanding relevant queries	not for all	yes for all
Invariant to monotonic transformation	no	yes