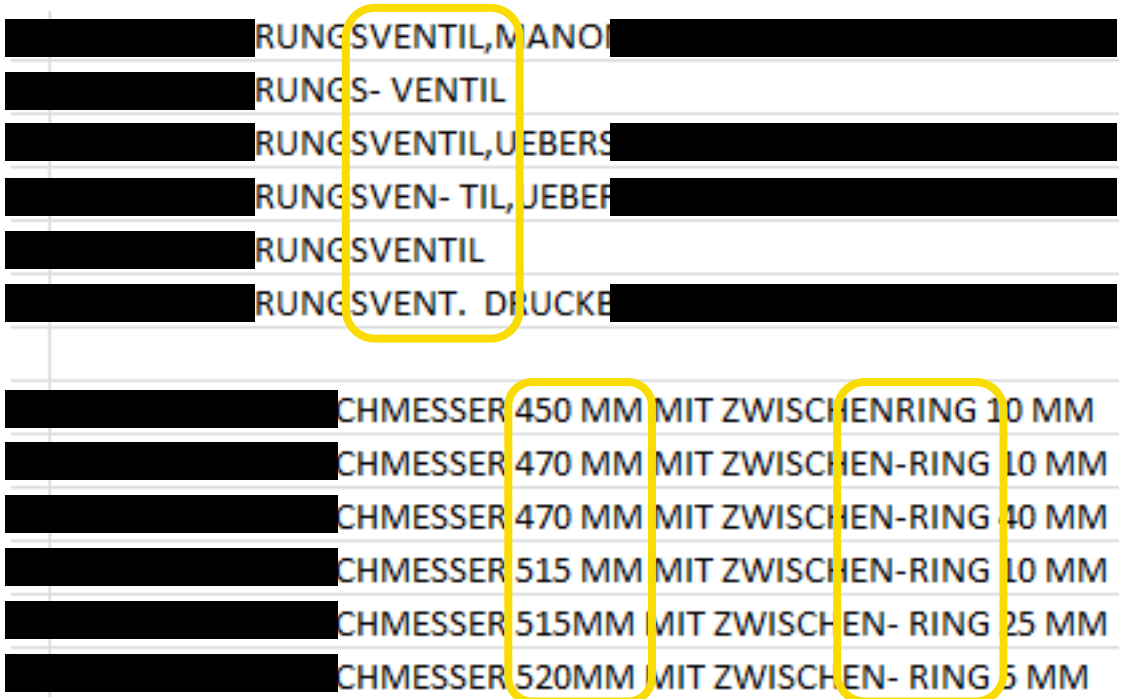
Verjüngungskur für sprachliche Altdaten

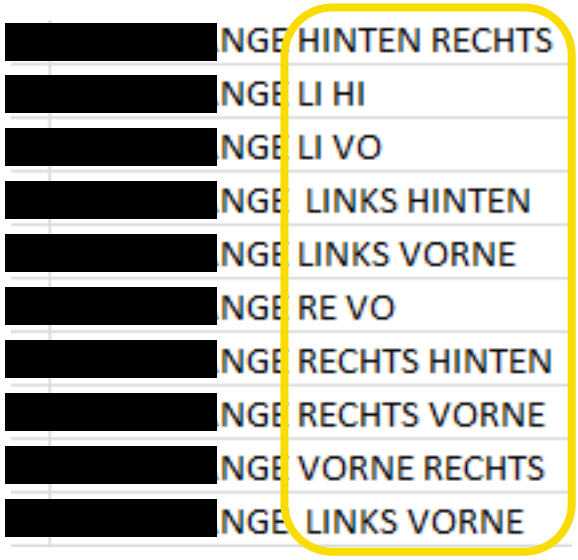
tekom-Jahrestagung 2017
24. bis 26. Oktober, Stuttgart

Axel Theofilidis
axel.theofilidis@iailc.de
IAI Linguistic Content AG, Saarbrücken

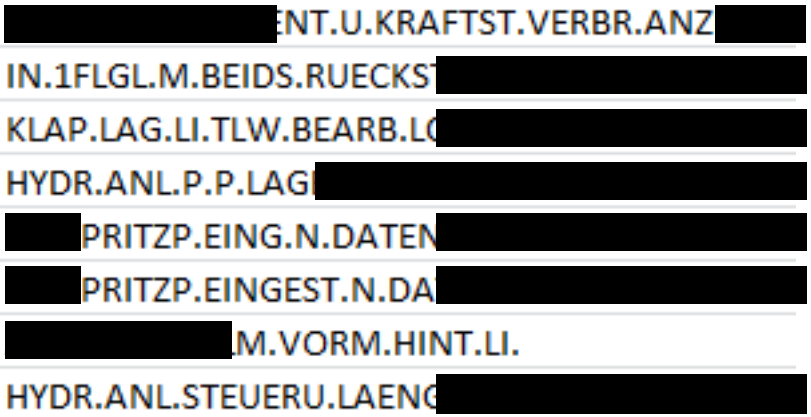
Kommt Ihnen das bekannt vor?



Oder das?


[redacted] NGE HINTEN RECHTS
[redacted] NGE LI HI
[redacted] NGE LI VO
[redacted] NGE LINKS HINTEN
[redacted] NGE LINKS VORNE
[redacted] NGE RE VO
[redacted] NGE RECHTS HINTEN
[redacted] NGE RECHTS VORNE
[redacted] NGE VORNE RECHTS
[redacted] NGE LINKS VORNE

Und was sagt Ihnen das?



Eher wenig?!

Dann sind Sie hier richtig!

Übersicht

- Die Beharrlichkeit sprachlicher Altdaten
- Was sprachliche Altdaten „auszeichnet“
- Ist die Bereinigung sprachlicher Altdaten machbar?
- Die stufenweise Bereinigung sprachlicher Altdaten
- Was die Technische Redaktion zum Gelingen beiträgt
- Fazit und Ausblick

Die Beharrlichkeit sprachlicher Altdaten

Die Beharrlichkeit sprachlicher Altdaten

Erfassung von Produktinformationen
seit Einführung der ersten EDV-Systeme

- Rasante Zunahme des Datenvolumens …
- … bei beschleunigter Entwicklung der IT-Technologie
- … und zunehmender Durchdringung der Betriebsabläufe
  (Arbeitspositionen, Kataloge, Vertrieb, Logistik, Shop, …)

⇒ Notwendigkeit der **Datenstabilität**

Die Beharrlichkeit sprachlicher Altdaten

Notwendigkeit der **Datenstabilität**

- EDV-Systeme werden migriert,
- die Daten selbst bleiben aber unangetastet,
- althergebrachte Eingabeverfahren werden beibehalten!

⇒ Es entstehen massenhaft Daten, die den heutigen Anforderungen an die sprachliche Qualität nicht genügen …

⇒ … und der **Datenkompatibilität** zuwiderlaufen!

Was sprachliche Altdaten „auszeichnet“

Was sprachliche Altdaten „auszeichnet“

Eingeschränkte Eingabemöglichkeiten
in älteren EDV-Systemen:

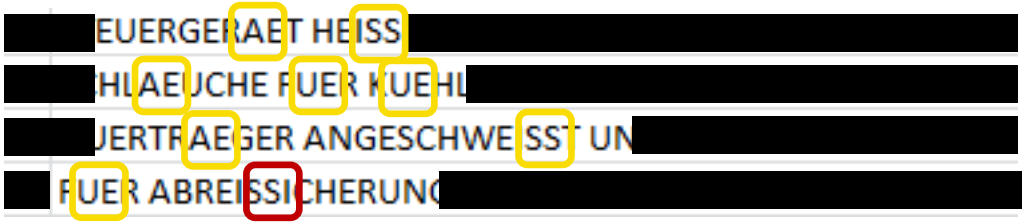
- limitierte Eingabefelder (30, 40, 80 Zeichen)
- limitierter Zeichensatz (typischerweise nur Großbuchstaben)

⇒ niedrige Hemmschwelle für
sprachliche Nachlässigkeiten jeglicher Art

⇒ … zumal es oftmals an der Technologie mangelt,
redaktionelle Vorgaben durchzusetzen!

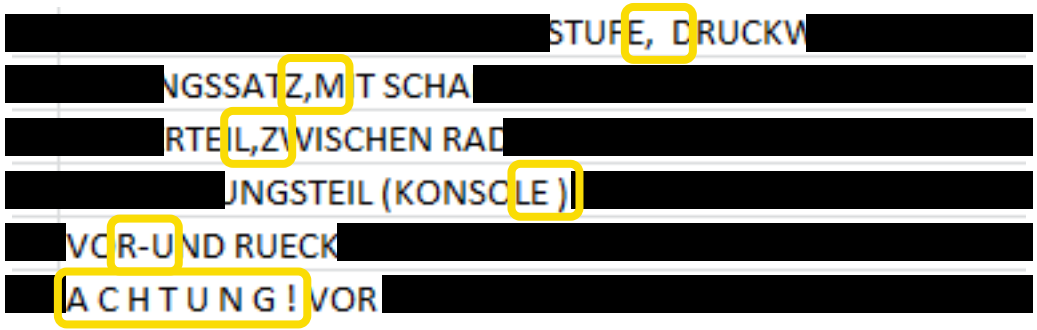
Was sprachliche Altdaten „auszeichnet“

- durchgängige Großschreibung (Versalschreibung)
- keine Umlaute, kein *ß*



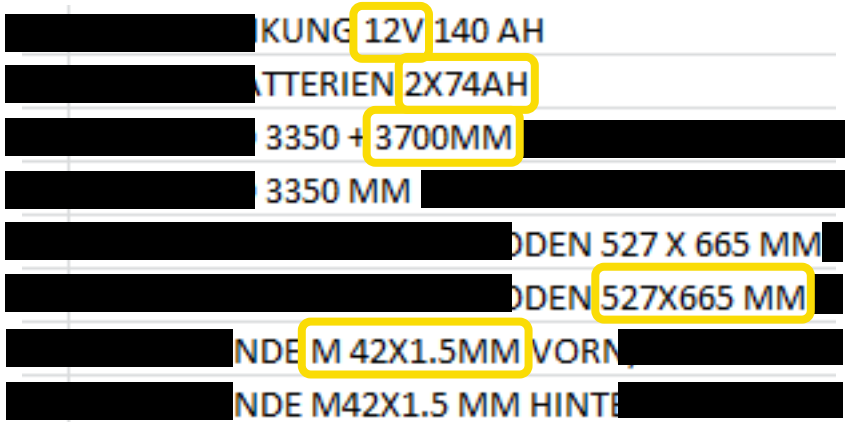
Was sprachliche Altdaten „auszeichnet“

- willkürliche Setzung (und Nicht-Setzung) von Leerzeichen
- Sperrsatz



Was sprachliche Altdaten „auszeichnet“

- unsystematische Eingabe spezieller Datentypen
  (Maßangaben, Schraubenbezeichnungen, Teilenummern usw.)

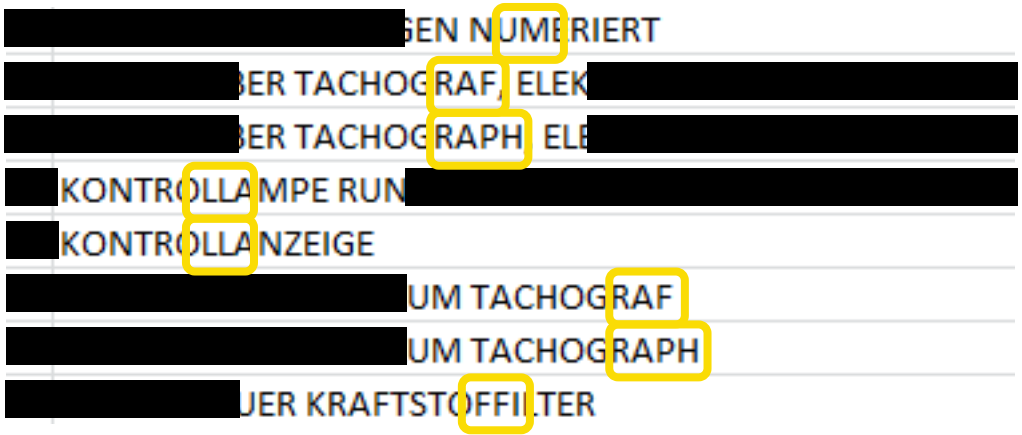


Was sprachliche Altdaten „auszeichnet“

- massenhaft Abkürzungen
  (auch Spontanabkürzungen, Binnenabkürzungen)

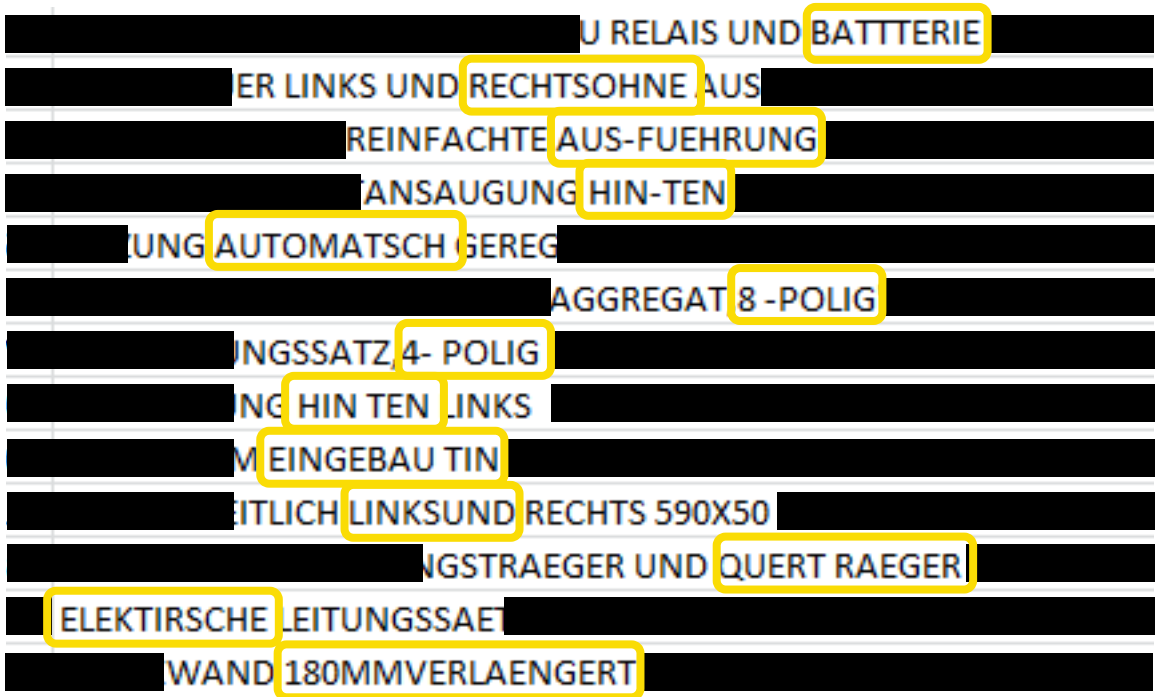
Was sprachliche Altdaten „auszeichnet“

- alte Rechtschreibung und orthografische Varianten



Was sprachliche Altdaten „auszeichnet“

- vielfältige Falschschreibungen (insbesondere falsche Zusammen- oder Getrenntschreibungen und Bindestrichsetzungen)



Ist die Bereinigung sprachlicher Altdaten machbar?

Ist die Bereinigung sprachlicher Altdaten machbar?

Eine manuelle Bereinigung ist
ökonomisch und intellektuell kaum leistbar!

- Redakteure müssten Spontanabkürzungen auflösen,
- sprachliche Fehler ausmerzen,
- terminologische Inkonsistenzen ausräumen,
- jeweils konsistent über den Gesamtdatenbestand!

Bei ± 1 Minute pro Kurztext von 50 bis 80 Zeichen

→ ~ 200 Personentage für 100.000 Datensätze

… optimistisch geschätzt!

Ist die Bereinigung sprachlicher Altdaten machbar?

Zielsetzung

- signifikant reduzierte Aufwände
- bei konstanter Korrekturleistung
- durch Einsatz von weitestgehend automatischen Verfahren

Was bedarf es hierfür mindestens?

- Einfache **Programmroutinen**, die spezielle Zeichenkettenersetzungen ausführen, aber auch …
- **linguistisch intelligente Sprachverarbeitungssoftware** zur automatischen Etablierung der korrekten Rechtschreibung und der terminologischen Konsistenz

Ist die Bereinigung sprachlicher Altdaten machbar?

Anforderungen an eine Sprachverarbeitungssoftware

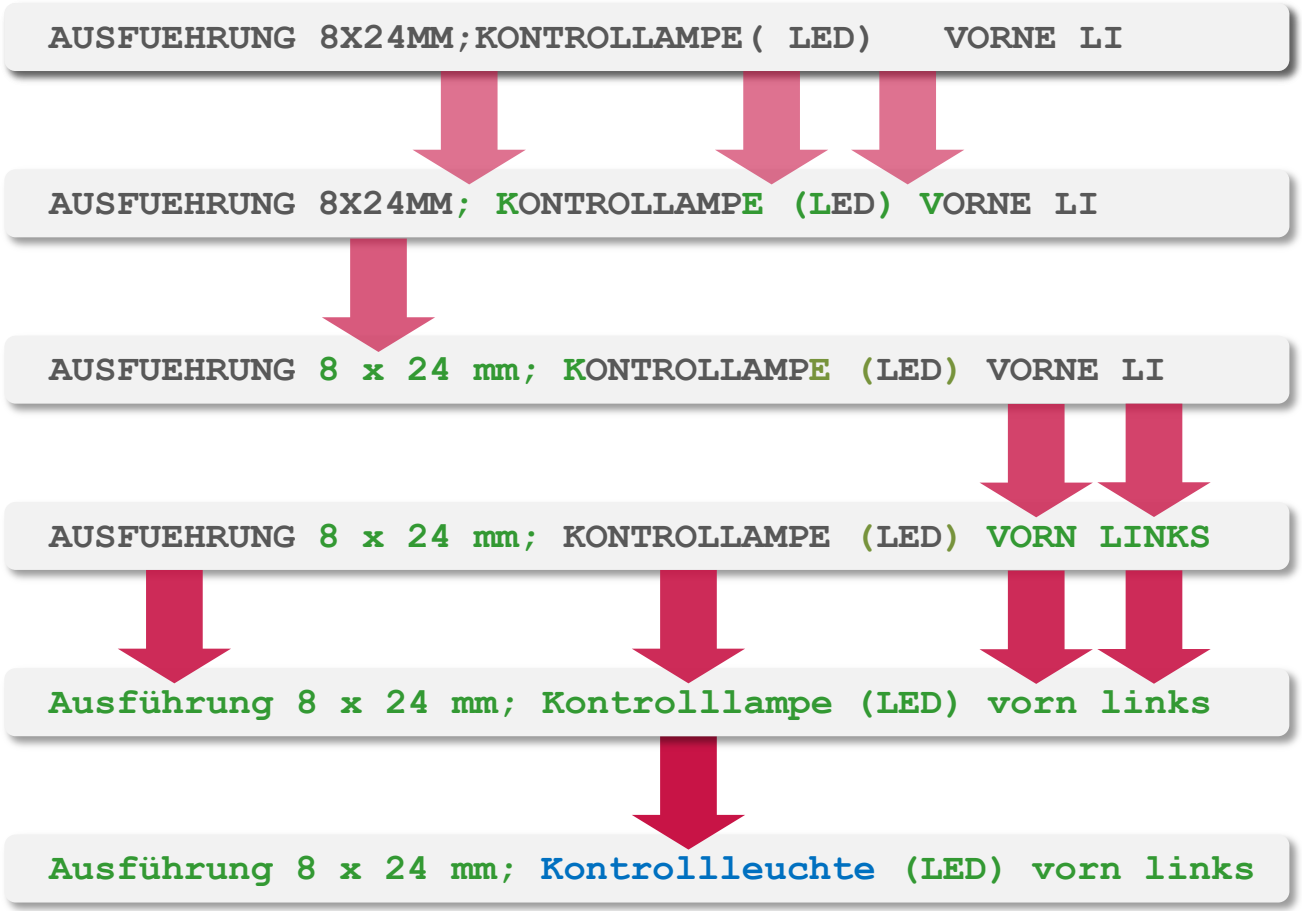
- umfassende **Abdeckung des relevanten Vokabulars**
- Funktion zur **Umwandlung von Versalschreibung** in normale Groß- und Kleinschreibung
- automatische **Korrektur systematischer Falschschreibungen** (z.B. ausgeschriebene Umlaute, alte Rechtschreibung)
- gewichtete Vorschläge für die **Korrektur von Tippfehlern**
- **Auflösung von Mehrdeutigkeiten** basierend auf einer grammatischen Analyse
- Idealerweise **Abgleich mit Unternehmensterminologie**

Ist die Bereinigung sprachlicher Altdaten machbar?

⇒ Mit einer entsprechenden Sprachverarbeitungssoftware kann die Bereinigung sprachlicher Altdaten in einem gestuften Verfahren angegangen werden.

Die stufenweise Bereinigung sprachlicher Altdaten

Die stufenweise Bereinigung sprachlicher Altdaten



- Leerzeichen normalisieren
- Datentypen standardisieren
- Zeichenketten ersetzen
- Rechtschreibung korrigieren
- Terminologie konsolidieren

Die stufenweise Bereinigung sprachlicher Altdaten

Normalisierung der **Leerzeichensetzung**
durch einfache Ersetzungsroutinen

- Mehrfachleerzeichen auf einzelne Leerzeichen reduzieren
- Leerzeichen am Zeilenanfang oder -ende löschen
- Leerzeichen im Kontext von Satzzeichen setzen bzw. löschen (vor/nach Klammer, Doppelpunkt, Semikolon, Komma)
  …

→ Berücksichtigung **besonderer Datenkonstrukte**:
gegliederte Nummern (13 45.00 04), Dezimalzahlen (20,5MM), Ordinalzahlen (1.STUFE), Verhältnisangaben (30:50)!

Die stufenweise Bereinigung sprachlicher Altdaten

Standardisierung der **Darstellung spezieller Datentypen** durch einfache (Muster-basierte) Ersetzungsroutinen

- Datumsangaben (24.9.2017)
- Maßangaben (30X50X200MM)
- Schraubenbezeichnungen (M42X1.5)

  …

→ Setzung von **Leerzeichen bzw. Festabständen**

→ Darstellung von **Sonderzeichen**
(z.B. **x** versus × als Mal-Zeichen)

Die stufenweise Bereinigung sprachlicher Altdaten

Nach Normalisierung der Leerzeichensetzung und Standardisierung der Darstellung spezieller Datentypen **erweisen sich erste Texte als Dubletten!**

⇒ Ein Anteil von **3% oder mehr Textdubletten** ist an dieser Stelle keine Seltenheit!

Die stufenweise Bereinigung sprachlicher Altdaten

Analyse des Zeichenketteninventars
durch Datenauswertungsroutinen und Wortanalysen

- Welche Zeichenketten
  kommen im Datenbestand wie oft vor?

144557	,
45645	/
41006	MIT
33004	;
28938	UND
24761	FUER
24169	:
21800	LINKS
21735	RECHTS
19578	AB
17477	BEI
16734	IN
14779	(
14676	)
12972	VON
12610	MM
12099	CODE
10892	EINGEBAUT
10329	TEILNUMMER

Die stufenweise Bereinigung sprachlicher Altdaten

Analyse des Zeichenketteninventars
durch Datenauswertungsroutinen und Wortanalysen

- Welche Zeichenketten kommen im Datenbestand wie oft vor?
- Welche weiteren speziellen Datentypen lassen sich ablesen?

2092	**1X**
849	**24V**
827	**12V**
727	**4X2**
535	**2X**
429	**4X4**
407	**V6**
361	**6X2**
302	**5T**
272	**6X4**
218	**8X4**
188	**4X**
180	**400L**
144	**300L**
121	**A31**
120	**3X**
103	**1X322**
81	**28V**
79	**1B**

Die stufenweise Bereinigung sprachlicher Altdaten

Analyse des Zeichenketteninventars
durch Datenauswertungsroutinen und Wortanalysen

- Welche Zeichenketten kommen im Datenbestand wie oft vor?
- Welche weiteren speziellen Datentypen lassen sich ablesen?
- Welche Abkürzungen sind im Datenbestand belegt?

1564	FA.
659	BZW.
385	NR.
311	FUSSN.
128	BEF.
113	U.
102	ELEKTR.
93	MIN.
68	NO.7
43	MECH.
39	F.
31	ERN.
27	LI.
24	M.P.H.
22	L.
21	BZW.BIS
20	BZW.AB
20	DURCHM.
18	S.

Die stufenweise Bereinigung sprachlicher Altdaten

Analyse des Zeichenketteninventars
durch Datenauswertungsroutinen und Wortanalysen

- Welche Zeichenketten kommen im Datenbestand wie oft vor?
- Welche weiteren speziellen Datentypen lassen sich ablesen?
- Welche Abkürzungen sind im Datenbestand belegt?
- Und welche unbekannten Wörter oder Falschschreibungen kommen eigentlich vor?

20	AUSSNAHME
20	VERSCHLUSSCHRAU
20	VORDER
19	FUSSTUETZE
19	AUSNAME
16	STUEZWINKEL
15	AUSNHAME
14	AUSRALIEN
13	AUSERDEM
12	OESTEREICH
11	UEBERSETUNG
10	VONMITTE
10	AUSFUERUNG
9	HAEUSE
9	EINGEBAUTIN
9	NUMMERIC
8	LOESCEN
7	ELECTRIC
7	INSIEHE

Die stufenweise Bereinigung sprachlicher Altdaten

Bereitstellung der Ergebnisse
der Zeichenkettenanalyse als einfaches Tabellenformat zur Bearbeitung seitens der Redaktion

	ubekannter String	Korrekturstring	Typ	Bemerkungen
26	[redacted]	[redacted]	name	[redacted]
20	FUSSTUETZE	Fußstütze	wrong	
19	AUSNAME	Ausnahme	wrong	
13	BEICODE	bei Code	split	
13	[redacted]		name	
10	BRENSE	Bremse	wrong	
10	EINGEBAUTIN	eingebaut in	split	
10	AUSSNAHME	Ausnahme	wrong	
5	LIGHTS	Lights	foreign	

Die stufenweise Bereinigung sprachlicher Altdaten

Generierung von Ersetzungsregeln
aus den bearbeiteten Tabellen der Zeichenkettenanalyse

→ Ausführung der Ersetzungen über den Gesamtdatenbestand

- **Expansion von Abkürzungen**, ggf. auch unter Berücksichtigung vorhandener Abkürzungslisten, z.B.

LI	→ LINKS
HI	→ HINTEN
BEZ	→ BEZEICHNUNG
REP.SATZ	→ REPARATURSATZ

Die stufenweise Bereinigung sprachlicher Altdaten

- **Auftrennung falscher Zusammenschreibungen**, z.B.

  BEICODE → BEI CODE
  EINGEBAUTIN → EINGEBAUT IN

- **Auswahl bei Wortvarianten**, z.B.

  WAAGERECHT → WAAGRECHT
  VORNE → VORN

- **Korrektur von unbekannten Wörtern** ohne eindeutigem Korrekturvorschlag, z.B.

  AUSNAME → AUSNAHME;AUSNAHMEN
  BRENSE → BRENNE;BREMSE

Die stufenweise Bereinigung sprachlicher Altdaten

- **Bekanntmachung unbekannter „Wörter“**, die nicht – oder auf bestimmte Weise – korrigiert werden sollen (Akronyme, Codes, Firmennamen usw.), z.B.

  DFT
  Z97G67
  TOTAL

  → ggf. **Aufnahme in ein Spezialwörterbuch** der Sprachverarbeitungskomponente

⇒ Der Datenbestand ist „reif“ für die linguistische Analyse und Rechtschreibkorrektur!

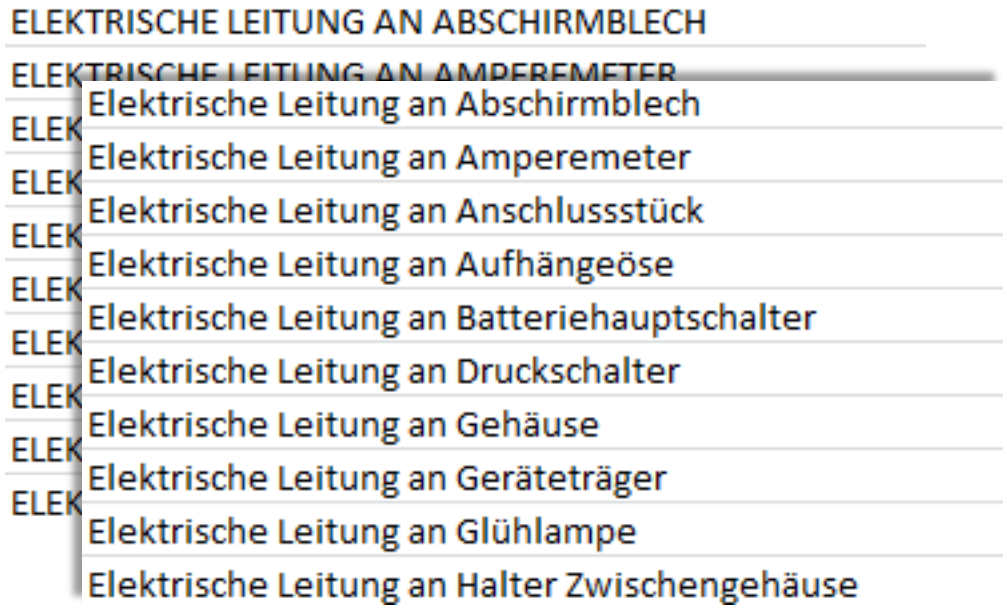
Die stufenweise Bereinigung sprachlicher Altdaten

Rechtschreibkorrektur basierend auf einer linguistisch fundierten Sprachverarbeitungskomponente

- **95%** und mehr der in Versalschreibung vorliegenden Texte können vollständig in **korrekte Groß- und Kleinschreibung** umgewandelt werden.
- Ausgeschriebene Umlaute und sonstige **Falschschreibungen** sind **umfassend und konsistent korrigiert**.

Die stufenweise Bereinigung sprachlicher Altdaten

Rechtschreibkorrektur basierend auf einer linguistisch fundierten Sprachverarbeitungskomponente



Die stufenweise Bereinigung sprachlicher Altdaten

Nur wenige Texte enthalten jetzt noch
einzelne Wörter, die nicht korrigiert wurden.

→ Grund hierfür: Wortmehrdeutigkeiten, die im gegebenen Kontext nicht aufgelöst werden konnten, z.B.

MASSE Ladedose → Masse / Maße ?

⇒ **Spezielle Auszeichnung** solcher Fälle zur Unterstützung einer gezielten manuellen Nachbereinigung

⇒ Auch **datenspezifische Vereindeutigungsregeln** sind denkbar!

Die stufenweise Bereinigung sprachlicher Altdaten

Kann die Rechtschreibkorrektur auch
über das Ziel hinausschießen?

→ Fehlanalysen und somit auch Fehlkorrekturen sind vereinzelt möglich, z.B.

ANSCHLUSS AN TRAEGER PUMPE → Anschluss an träger Pumpe

⇒ Erfassung aller Wortmehrdeutigkeiten

⇒ Implementierung **datenspezifischer Vereindeutigungsregeln**

Die stufenweise Bereinigung sprachlicher Altdaten

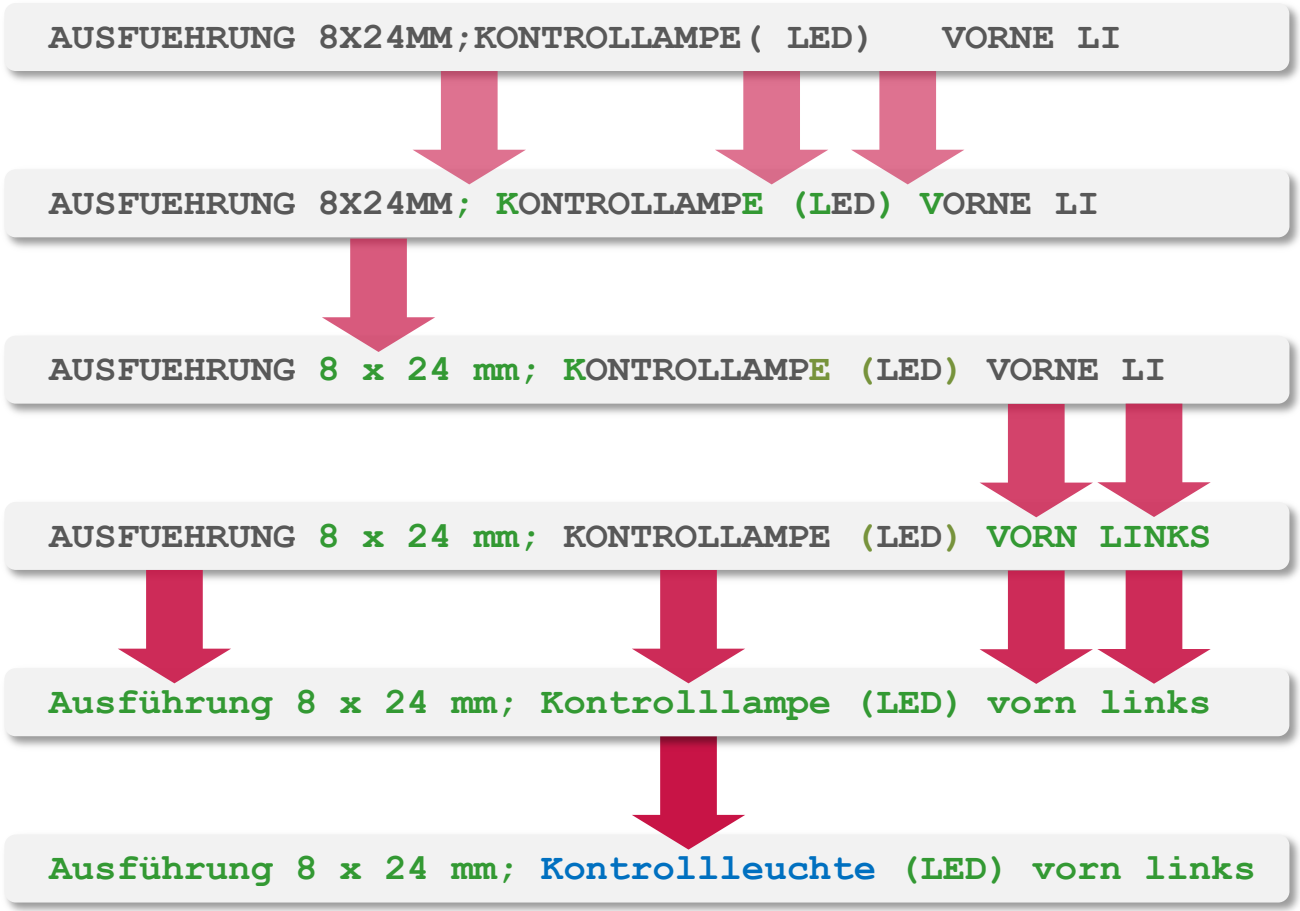
Terminologiekonsolidierung basierend auf einem Abgleich mit der Unternehmensterminologie

- Erkennung von verbotenen Termen und Termvarianten
- Ersetzung durch Vorzugsbenennungen, z.B.

  Kontrolllampe $\rightarrow$ Kontrollleuchte
- … oder zumindest Auszeichnung als terminologische Problemfälle

$\Rightarrow$ Die stufenweise Bereinigung sprachlicher Altdaten auf der Grundlage automatischer Verfahren ist damit fürs Erste abgeschlossen!

Die stufenweise Bereinigung sprachlicher Altdaten



- Leerzeichen normalisieren
- Datentypen standardisieren
- Zeichenketten ersetzen
- Rechtschreibung korrigieren
- Terminologie konsolidieren

Was die Technische Redaktion zum Gelingen beiträgt

Was die Technische Redaktion zum Gelingen beiträgt

Die **Technische Redaktion unterstützt maßgeblich** den Prozess der Bereinigung sprachlicher Altdaten.

- Abstimmung mit den redaktionellen Vorgaben
- Anpassung an die spezifischen Datengegebenheiten

⇒ Erhöhung der Bereinigungsquote

⇒ geringere Aufwände für manuelle Nachbereinigung

⇒ Vermeidung von Fehlkorrekturen

Was die Technische Redaktion zum Gelingen beiträgt

Bereitstellung von Wissensquellen

- Redaktionsleitfaden
  → Muster-Spezifikationen für spezielle Datentypen
- Abkürzungslisten
  → Expansion der Langformen
- Terminologie
  → Einsetzung der Vorzugsbenennungen
- Marketingbenennungen, Produktnamen, Zulieferfirmen …
  → Vermeidung von Fehlkorrekturen

Was die Technische Redaktion zum Gelingen beiträgt

Iterative Bewertung von Zwischenergebnissen,
insbesondere der Analyse des Zeichenketteninventars

- Expansion von Spontan- und Binnenabkürzungen
- Festlegungen zu Tippfehlerkorrekturen und Wortvarianten
- Klassifikation unbekannter Wörter als Spezialvokabular
- Stichprobenprüfungen

Was die Technische Redaktion zum Gelingen beiträgt

Manuelle Nachbereinigung

- Sichtung speziell ausgezeichneter Textstellen

→ nicht korrigierte Wörter (wegen Wortmehrdeutigkeiten)

→ terminologische Problemfälle

⇒ Auch die letzten verbleibenden Problemfälle sind intellektuell bewertet und korrigiert!

Fazit und Ausblick

Fazit und Ausblick

Die über viele Jahre (Jahrzehnte) aufgeschobene Bereinigung sprachlicher Altdaten ist machbar!

Fazit und Ausblick

Ein Bereinigungsgrad von 95% und mehr ist erreichbar

- mit einigen grundlegenden Ersetzungsroutinen
- … und dem Einsatz einer hochwertigen Sprachverarbeitungskomponente.

Ein höherer Bereinigungsgrad wird durch die **Abstimmung des Bereinigungsprozess auf die unternehmensspezifischen Gegebenheiten** erzielt.

→ Berücksichtigung redaktioneller Vorgaben und Referenzdaten

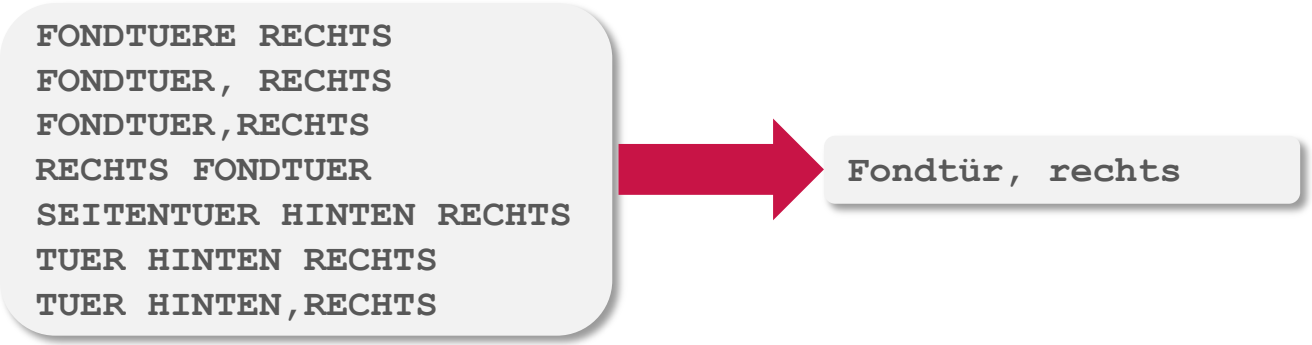
→ Einbeziehung der Technischen Redaktion

Fazit und Ausblick

Variantenermittlung

auf der Grundlage der bereinigten Daten

- Benennungsvarianten
- Wortstellungsvarianten
- Formulierungsvarianten



Fazit und Ausblick

→ Eliminierung unnötiger Textvarianten

→ Reduktion der informationellen Redundanz im Gesamtdatenbestand

⇒ Fit für den effektiven Einsatz sprachtechnologischer Anwendungen wie die eines TM oder AM

Fazit und Ausblick

⇒ **Fit für die Zukunft!**

Vielen Dank für Ihre Aufmerksamkeit.