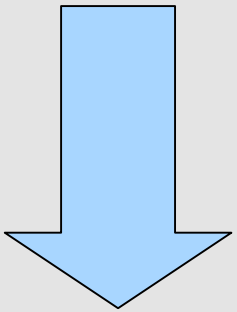
# Eigennamenerkennung mit großen lexikalischen Ressourcen

*Jörg Didakowski*
*BBAW*

# Was sind Eigennamen?

Sprachliche Zeichen (Eigennamen)



das Bezeichnete (Named Entity)

# Was sind Eigennamen?

Eigennamen bezeichnen ein Ding als Ganzes, ohne ihm bestimmte Eigenschaften zuzuschreiben.

Sie identifizieren Objekte innerhalb gekennzeichneter Klassen.

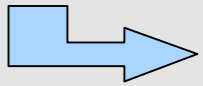
# Was sind Eigennamen?

- Organisationen
- Personen
- Orte
- Produkte
- Zeitangaben
- Datumsangaben
- Währungsangaben

# Was sind Eigennamen?

- Organisationen
- Personen
- Orte
- Produkte
- Zeitangaben
- Datumsangaben
- Währungsangaben

# Was sind Eigennamen?

- Organisationen
- Personen
- Orte
- Produkte
  -  Gattungsnamen oder Stoffnamen (Opel/Auto, Esso/Wasser)
- Zeitangaben
- Datumsangaben
- Währungsangaben

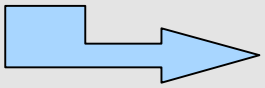
# Offene Klasse

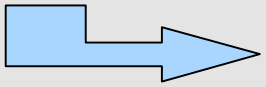
Eigennamen bilden eine Offene Klasse – täglich werden neue Namen erfunden

Jedes Wort des allgemeinsprachlichen Lexikons kann zum Eigennamen gemacht werden (Ihr Name war Dachziegel)

# Morphosyntaktische Besonderheiten

Eigennamen treten im Allgemeinen im Singular auf. Ausnahmen sind:

- **Plural bei Familiennamen** ➔ die Bachs
  
  Keine Umlautbildung (Bach/Bäche)

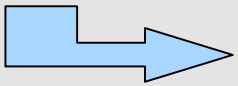
- Plural bei Ortsnamen ➔Niederlande, Alpen
  
  nur im Plural

- **Beim Quantifizieren** ➔ beide Deutschlands, die drei Marias

# Morphosyntaktische Besonderheiten

Eigennamen können die Form eines possessiven Genitivs einnehmen:

- Evas Auto, Brandenburgs Seen



artikellos

- Arturo Uis Aufstieg
- Der Aufstieg des Artuo Ui
- *Der Aufstieg des Arturo Uis

# Syntaktische Besonderheiten

Eigennamen treten Allgemein ohne Artikel auf:

- Sie liebt _ Frankreich.
- Sie liebt _ Berlin.
- Sie liebt _ Jürgen Klinsmann.

# Syntaktische Besonderheiten

Unter bestimmten Bedingungen kann systematisch ein bestimmter Artikel nötig sein:

- die Schweiz
- der Rhein
- das alte Berlin (Spezifische Kontextbedingungen).
- Wenn ein Organisationsname einen Appelativum enthält:
  die (Sozialdemokratische Partei Deutschland) SPD

# Syntaktische Besonderheiten

Bestimmter Artikel bei Personennamen, regionale Unterschiede :

- im Süddeutschen Sprachgebrauch besteht eine Tendenz des konnotationsfreien Gebrauchs von Personennamen mit Artikel.
- In anderen Regionen kann damit auf besonders bekannte Personen oder auch vertraute Personen verwiesen werden.

# Syntaktische Besonderheiten

Verwendung des unbestimmten Artikel signalisiert ganz besondere Interpretationsbedingungen:

- Er ist ein Cäsar → Appellativum

# Graphematische Besonderheiten

- Durchgehende Großschreibung am Wortanfang (Schwarzes Meer, Kap der Guten Hoffnung)

- das Schwarze Brett / das schwarze Brett

# Homographie

- Eigennamen können mehreren Klassen zugeordnet werden:
  Ford ➔ Person (Gerhard Ford)
  ➔ Firma (Ford Motors)
  ➔ Ort (Ford Michigan)
  (➔ Produktname (Automarke))
- Homographie zwischen Eigennamen und Appelativen (Fischer, Hirsch)
- Homographie zwischen Eigennamen und anderen Kategorien wegen Satzanfang (z.B. Als, Da, Kühn)

# interne Evidenz

- Dr. Bonobo
- Ich GmbH
- „Michael Schumacher“ ->im Lexikon
- Nicht Homographer Personenname (Sabine)
- Nicht Homographer Ortsname (Wustermark)

# externe Evidenz

Menschenbezeichner/Ortsbezeichner/Organisationsbezeichner als Apposition:

der Rennfahrer Schumacher
Schumacher, der Rennfahrer

der Rennstall Sauber
der Schweizer Sauber

# Koreferenz

- Sichere Eigennamen (externe/interne Evidenz)
- Unsichere Eigennamen (alle Kategorieabfolgen für einen potentiellen Eigennamen)
- Stützen von unsicheren Eigennamen wegen bestehender Koreferenz mit einem sicheren Eigennamen

# Ressourcen

TAGH-Morphologie (gewichteter Transduktor)

lemmatisiert und zerlegt Wortformen

Erkennungsrate bei neuen Zeitungstexten:
98,5% - 99,5%

# Ressourcen

- Nomenlexikon: 88.000 einfache und komplexe Stämme

- Eigennamen:
  160.000 geographische Eigennamen,
  65.000 Vornamen,
  240.000 Familiennamen

- Nomenthesaurus: 60.000 semantisch klassifizierte Nomen

- usw.

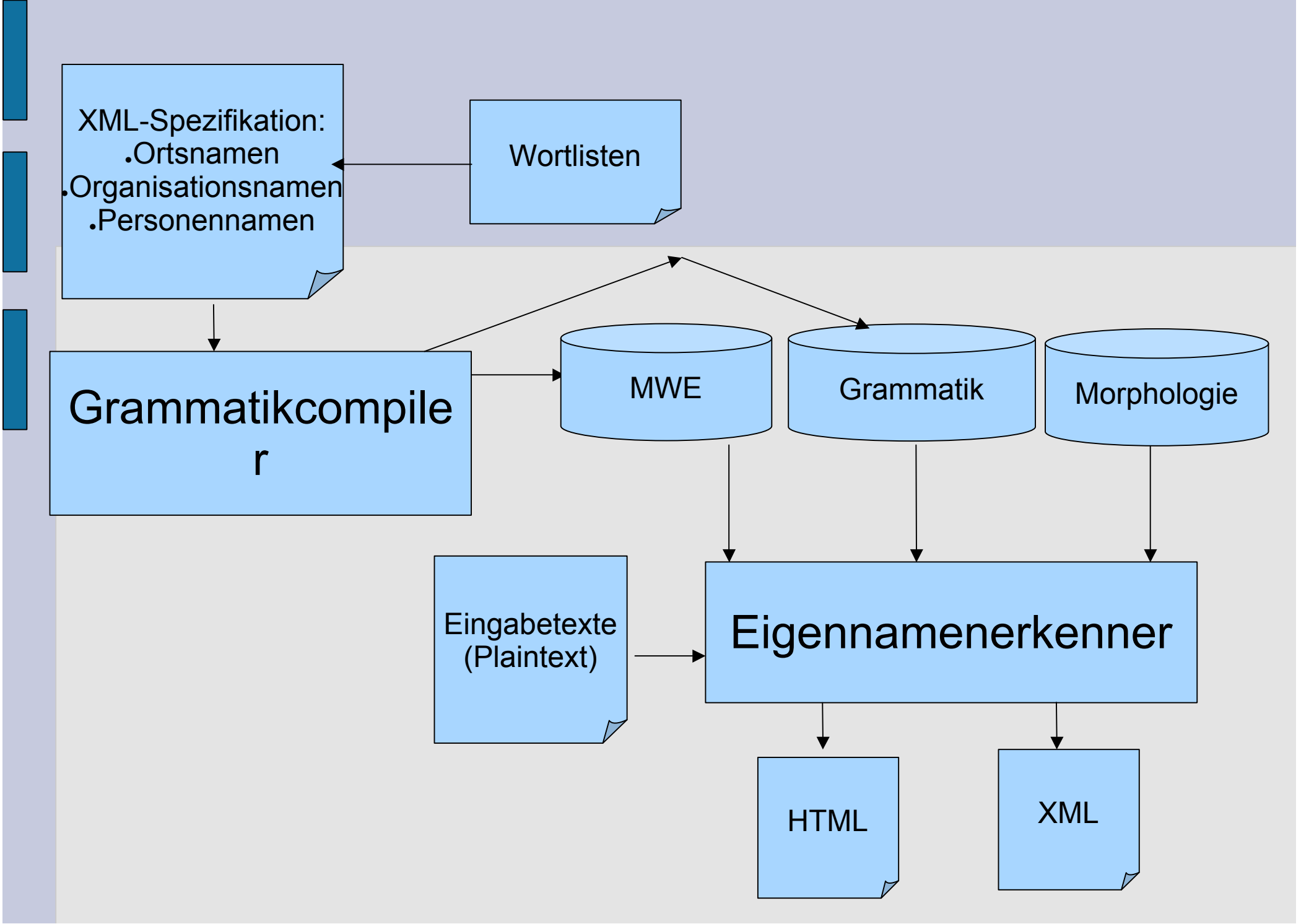
# Ressourcen

Listen von Mehrwortausdrücken

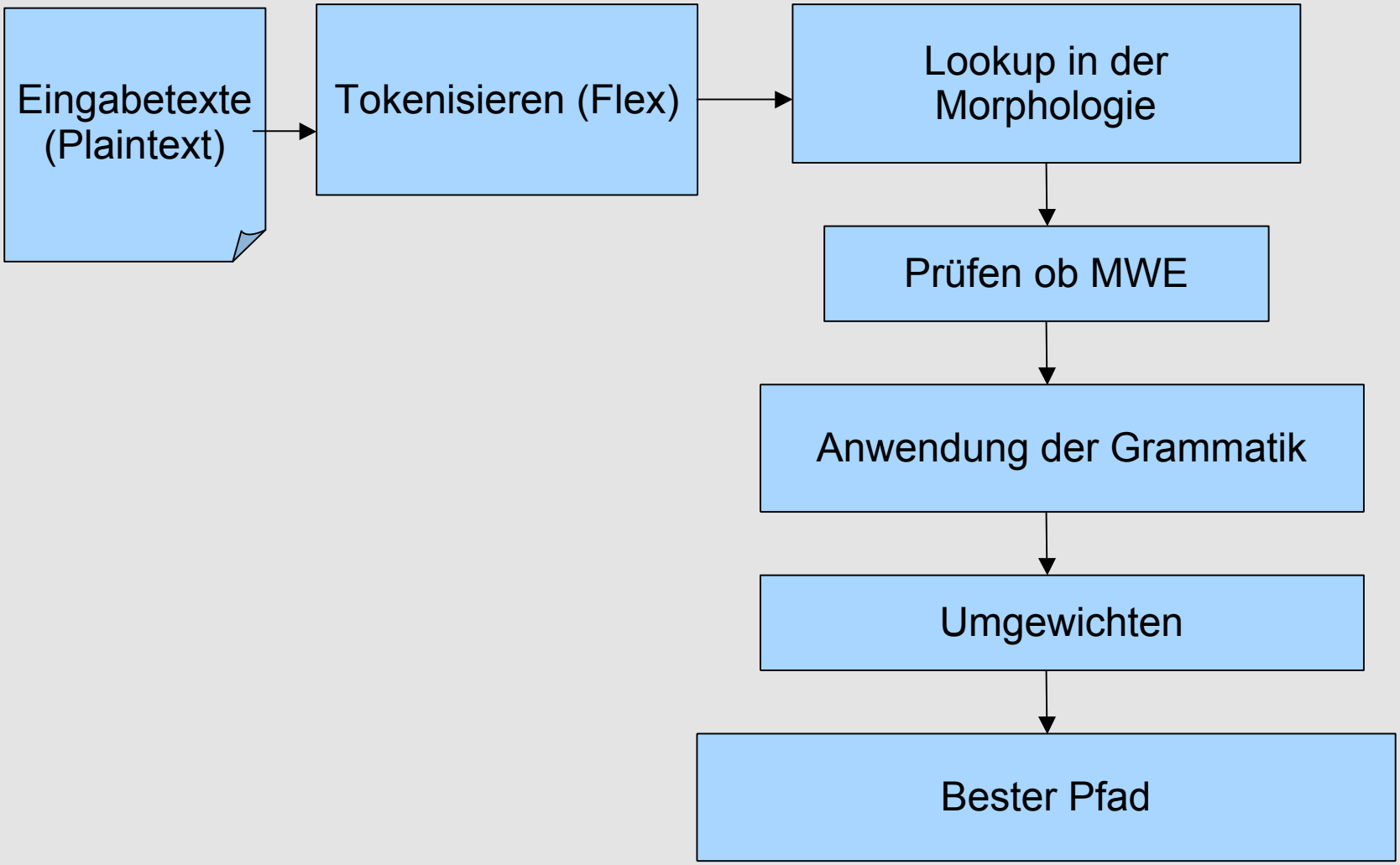
Ortsnamen (New York)
VIP Namen (Harald Schmidt)

# Das System (SynCoP)

- In C++ implementiert
- Basiert auf der Potsdam FSMlib

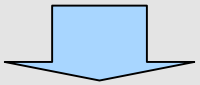

XML-Spezifikation:
•Ortsnamen
•Organisationsnamen
•Personennamen
Wortlisten
Grammatikcompiler
MWE
Grammatik
Morphologie
Eingabetexte
(Plaintext)
Eigennamenerkenner
HTML
XML

# Eigennamenerkenner



# Grammatikanwendung

Klammern und gewichten (Longest Match) von allen sicheren Eigennamen
Klammern und gewichten (Longest Match) von allen unsicheren Eigennamen

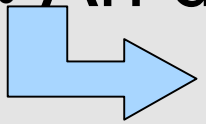


Umgewichten (sicher/unsicher)

Ermitteln der Besten Analyse
(Beste Pfad Suche)

- 10000 Tokens die Sekunde

- An der Grammatik wird noch entwickelt



Eine erschöpfende Evaluation wurde noch nicht durchgeführt.

Ende