

Bibliotheken und Mehrsprachigkeit aus der Perspektive von Automatisierung

Felix Sasaki

DFKI / FH Potsdam

felix.sasaki@dfki.de

Übersicht

- Was ist das mehrsprachige Web?
- Nutzen durch Bibliotheken
- Beiträge durch Bibliotheken

Inhalte

The screenshot displays a collage of W3C website pages in three languages: Korean, Greek, and Hindi. The Korean pages (top left and middle left) feature the W3C logo and text about the 'Video on the Web 2008' conference and the W3C Korea office. The Greek pages (top right and middle right) show the W3C Greece office information, including contact details and a list of members. The Hindi pages (bottom left and bottom right) focus on the W3C India office, providing details about the 'User Agent Accessibility Guidelines (UAAG) 2.0' and 'Mobile Web Best Practices' working groups. The bottom right page also includes a table of links for local and international members.

Interfaces

The image displays four overlapping screenshots of the W3C CSS Validation Service interface, each in a different language. The top-most screenshot is in Persian, followed by Korean, German, and Spanish. Each screenshot shows the service's logo, the title 'CSS Validation Service', and various navigation options. The Spanish screenshot is the most detailed, showing the 'Validar mediante URI' section with a text input field for the URI and a 'Check' button. At the bottom of the Spanish screenshot, there is a note about validating CSS in (X)HTML documents and a 'Donate' link. The footer of the Spanish screenshot includes the W3C logo, a copyright notice, and a 'Validator' button.

Zugriff auf Inhalte – Services zur Überbrückung von Sprachgrenzen



Google Übersetzer

Von:  Nach:

Geben Sie Text oder eine Website-Adresse ein oder [lassen Sie ein Dokument übersetzen](#).

Übersicht

- Was ist das mehrsprachige Web?
- Nutzen durch Bibliotheken
- Beiträge durch Bibliotheken

Nutzen I: Technische Infrastruktur

Unicode in Webtechnologien

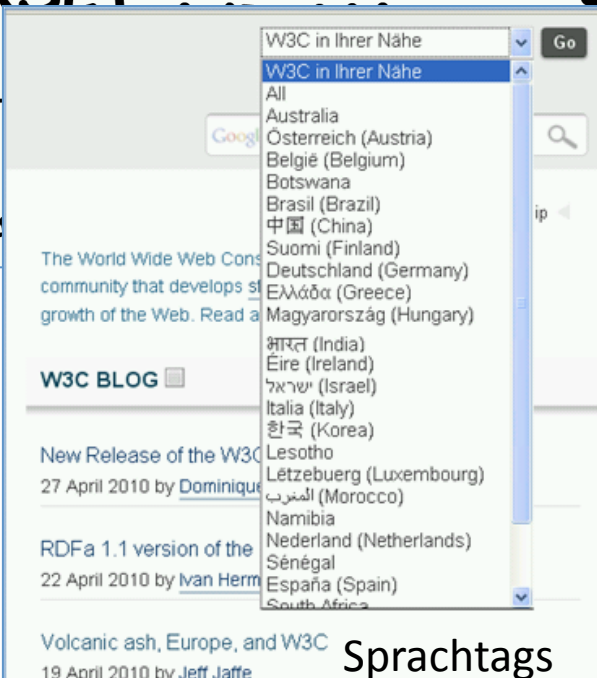
XPath OWL RDF SOAP

VoiceXML

XSLT SVG

Math

WA



W3C in Ihrer Nähe

All

Australia

Österreich (Austria)

België (Belgium)

Botswana

Brasil (Brazil)

中国 (China)

Suomi (Finland)

Deutschland (Germany)

Ελλάδα (Greece)

Magyarország (Hungary)

भारत (India)

Éire (Ireland)

ישראל (Israel)

Italia (Italy)

한국 (Korea)

Lesotho

Lëtzebuerg (Luxembourg)

المغرب (Morocco)

Namibia

Nederland (Netherlands)

Sénégal

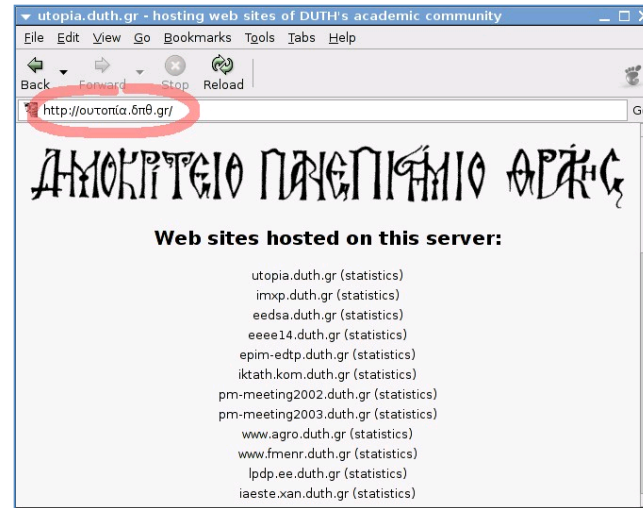
España (Spain)

South Africa

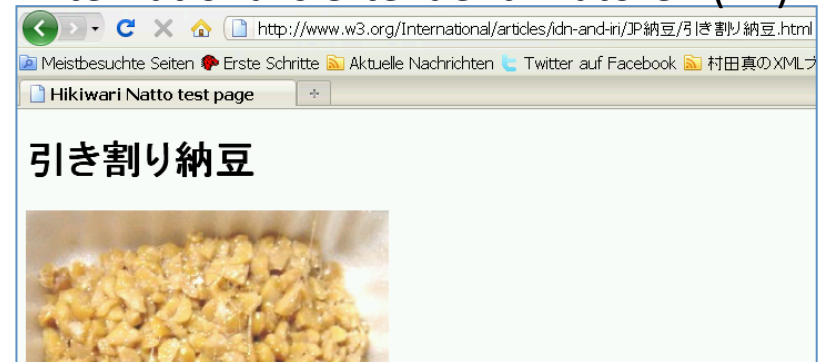
Sprachtags

wie "en", "en-us", "de", ...

Internationalisierte Domainnamen (IDN)



Internationalisierte Identifikatoren (IRI)



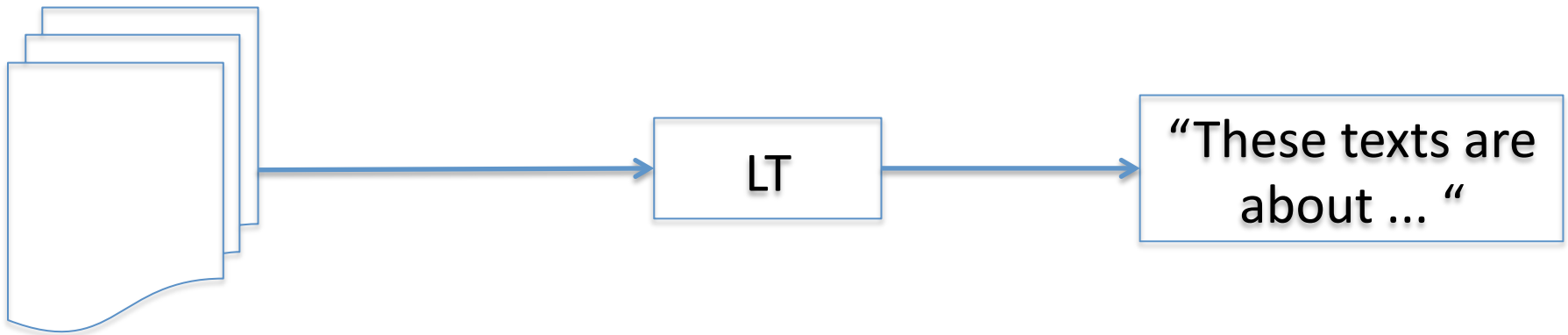
Nutzen II: automatisierte Informationsverarbeitung

- Suche über Sprachgrenzen hinweg
 - Automatische Übersetzung, Cross Language Information Retrieval
- Anreicherung mit (mehrsprachigen) Informationen
 - Verlinkung von Normdateien über Sprachgrenzen hinweg (DNB > VIAF > LC)

Nutzung von Sprachtechnologie („Language Technology“)

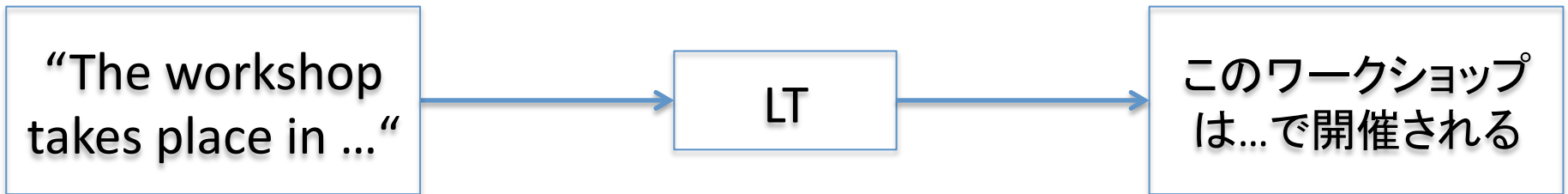
Sprachtechnologie: exemplarische Anwendungen

- Automatische Zusammenfassung



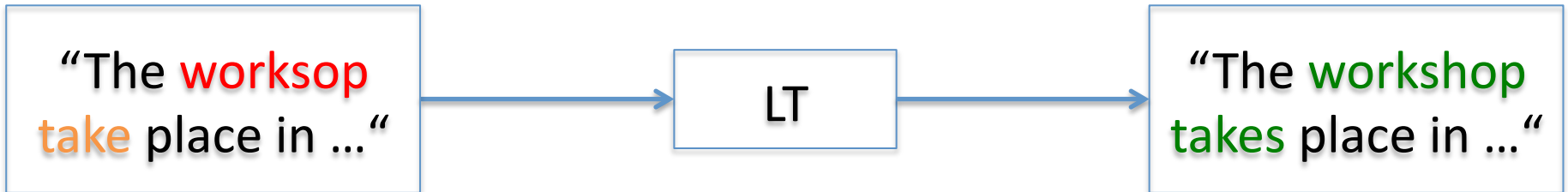
Sprachtechnologie: exemplarische Anwendungen

- Maschinelle Übersetzung



Sprachtechnologie: exemplarische Anwendungen

- Rechtschreib- / Grammatikkorrektur



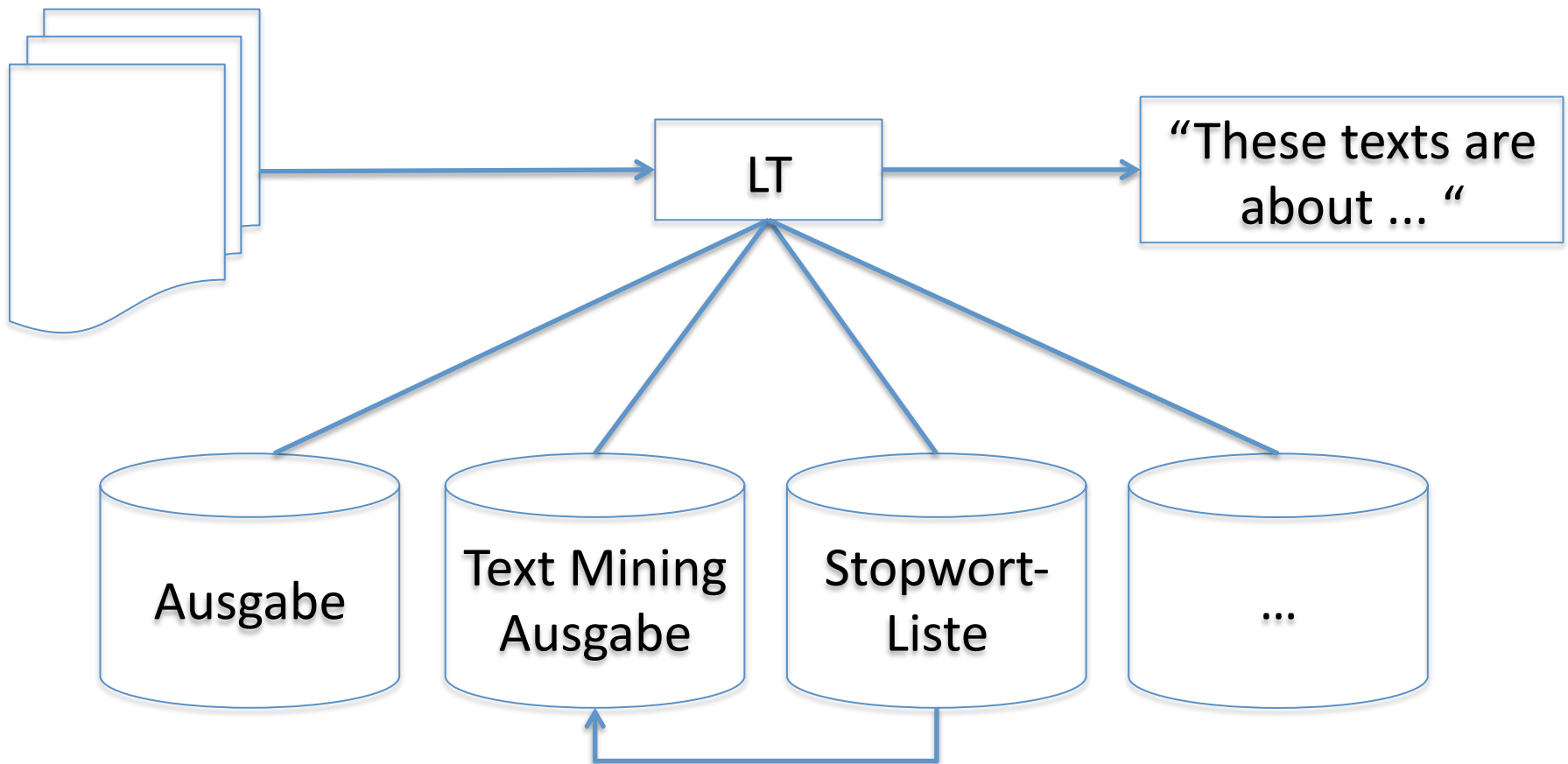
- Und viele weitere Anwendungen
 - Koreferenzauflösung, Diskursanalyse, Erkennung von Eigennamen, Generierung natürlicher Sprache, ...

Grundlagen:
Was haben alle diese
Anwendungen von
Sprachtechnologie gemeinsam?

Sie nutzen Ressourcen

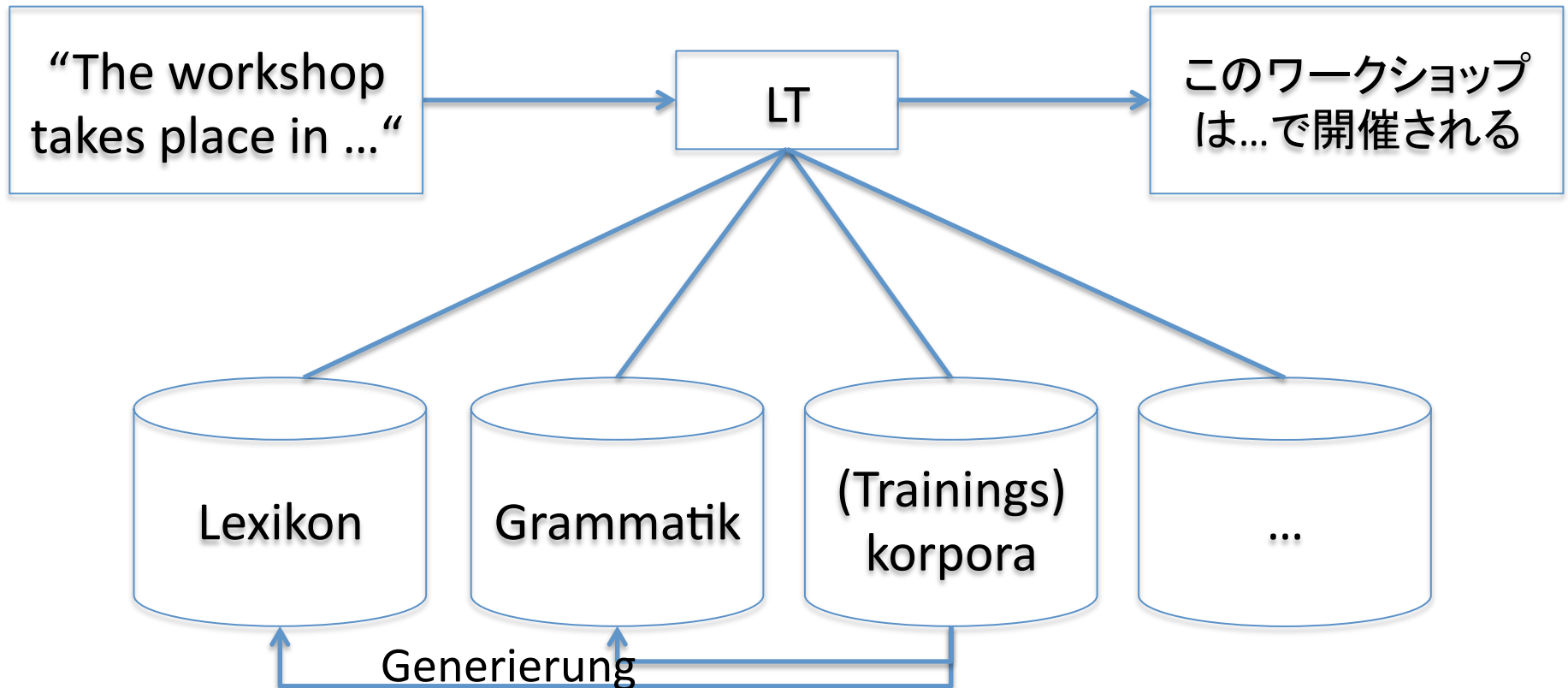
Ressourcen in der Sprachtechnologie

- Ressourcen für die Zusammenfassung



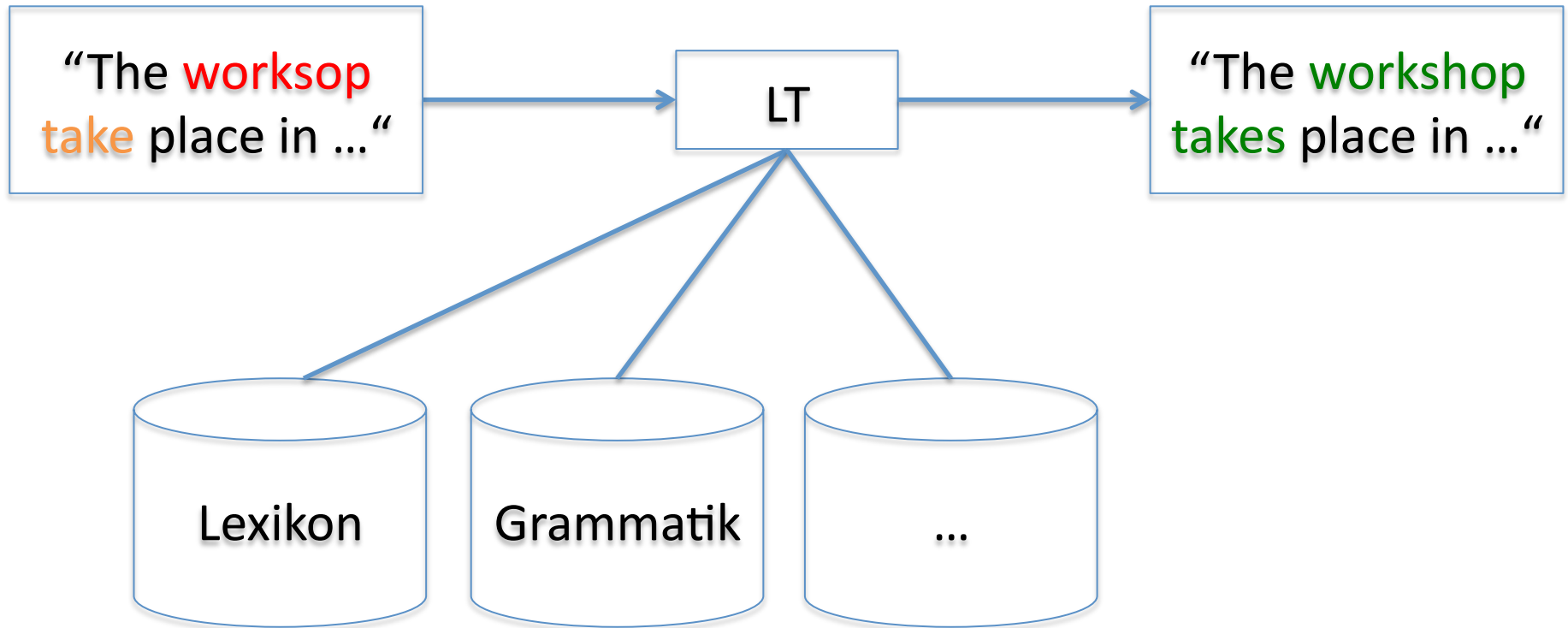
Ressourcen in der Sprachtechnologie

- Ressourcen für maschinelle Übersetzung



Ressourcen in der Sprachtechnologie

- Ressourcen für (Rechtschreib) Kontrolle



Vorläufige Zusammenfassung

- Bibliotheken nutzen im mehrsprachigen Web
 - Technische Infrastruktur
 - Automatische Methoden der mehrsprachigen Informationsverarbeitung
- Automatische mehrsprachige Informationsverarbeitung = Sprachtechnologie
 - Verschiedene Anwendungen (automat. Übersetzung, ...)
 - Immer: Nutzung von Ressourcen

Übersicht

- Was ist das mehrsprachige Web?
- Nutzen durch Bibliotheken
- Beiträge durch Bibliotheken

Vorläufige Zusammenfassung (Wiederholung und Frage)

- Bibliotheken nutzen im mehrsprachigen Web
 - Technische Infrastruktur
 - Automatische Methoden der mehrsprachigen Informationsverarbeitung
- Automatische mehrsprachige Informationsverarbeitung = Sprachtechnologie
 - Verschiedene Anwendungen (automat. Übersetzung, ...)
 - Immer: Nutzung von Ressourcen
 - Woher kommen die Ressourcen her?

Linguistische Ressourcen ...

- Linguistische Korpora
- Lexika
- „Ontologien“
- Einfache Tools
 - Morphologische Tagger
 - Syntaktische Parser
- Komplexe Tools
 - Erkennung von Eigennamen

... brauchen / umfassen Entitäten

- Linguistische Korpora
- Lexika
- „Ontologien“
- Einfache Tools
 - Morphologische Tagger
 - Syntaktische Parser
- Komplexe Tools
 - Erkennung von Eigennamen

Beispielnutzung von Entitäten (I)

- Anreicherung von Korpora mit Informationen aus Normdateien
 1. Eingabe (Ausschnitt): "Das von J. W. v. Goethe verfasste Buch ..."
 2. Semantische Verarbeitung: automatische Erkennung von Eigennamen
 3. Ergebnis: "Das von <pers pnd-link="118540238" viaf-link="24602065">J. W. v. Goethe</pers> verfasste Buch ..."

Beispielnutzung von Entitäten (II)

- Training und Evaluierung von Textkategorisierung, mit „getaggten“ digitalisierten Büchern
 - Eingabe: Thematisch manuell klassifizierte digitale Objekte
 - Verarbeitung z.B.
 - Training von LT-Tools die Texte kategorisieren
 - Evaluierung entsprechender Tools

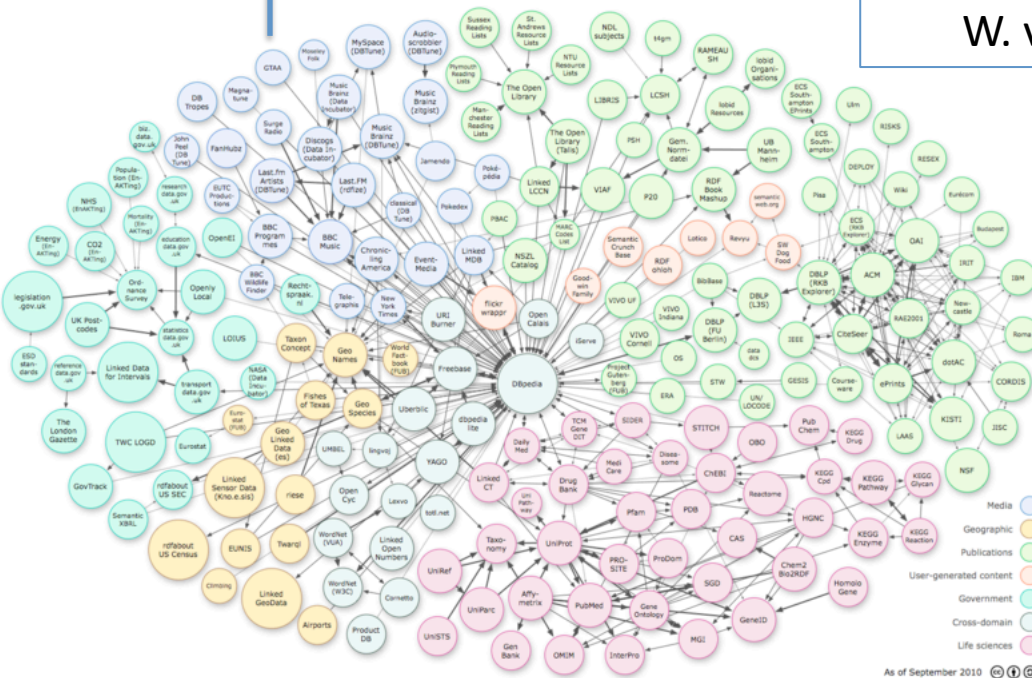
Sprachtechnologische Tools
verlassen sich in zunehmendem
Maße auf „externe Ressourcen“
wie „Linked Open Data“

1. Eingabe (Ausschnitt): "Das von J. W. v. Goethe verfasste Buch ..."

2. Semantische Verarbeitung:
automatische Erkennung von Eigennamen

3. Ergebnis: "Das von <pers pnd-link="118540238" viaf-link="24602065">J. W. v. Goethe</pers> verfasste Buch ..."

Datenquelle



Probleme

- Heterogene Qualität der Datenquellen
- Nachhaltigkeit
- Rechtliche Verlässlichkeit
- ...

Lösungen

- Bibliotheken haben das Potential ein Rückgrat für sprachtechnologische Verarbeitung im Web zu sein
- Einige Schwierigkeiten der müssen zur Einlösung dieses Potentials gelöst werden sein
- Stichwort „Datenheterogenität“
 - Syntaktisch
 - Semantisch

Syntaktische Heterogenität

Ausschnitt aus der Entitätsdefinition

<http://viaf.org/viaf/24602065/rdf.xml>

- Goethe,+J.+W.+v.
- Goethe,+Johann+Wolfgang+von
- VonGoethe,+Johann+Wolfgang
- Heterogenität für menschlichen Leser kein Problem
- Nutzung durch automatische Verarbeitung problematisch

Semantische Heterogenität

- Semantische Beziehungen für Nomen in WordNet
 - Hyperonym („Pudel“ ist Hyperonym von „Hund“),
Hyponym („Hund“ ist Hyponym von „Pudel“),
Holonym („Haus“ ist Holonym von „Fenster“),
Meronym („Fenster“ ist Meronym von „Haus“), ...
- Nicht immer abbildbar auf Beziehungen in bibliothekarischen Ressourcen
 - Was ist OB / UB / VB (Thesaurus), was ist
Oberklasse / Unterklasse (Klassifikation)

(Forschungs)agenda

- Bei der Bereitstellung von bibliothekarischen Ressourcen im (Semantic) Web – evaluieren:
 - Prinzipielle Eignung für ausgewählte Sprachtechnologische Anwendungen
 - Konkrete Probleme (mit dem Potential der Wiederverwendbarkeit) syntaktischer und semantischer Heterogenität
- Nachhaltigkeit der Bereitstellung
- Rechtliche Verlässlichkeit / freie Verfügbarkeit

(Forschungs)agenda

- Den kleinsten gemeinsamen Nenner finden
 - Syntaktische Heterogenität: schwierig
 - Semantische Heterogenität: möglich, vgl. „Europeana Data Model“
- ... und seine Anwendungen über Community-Grenzen hinweg sichern
 - Bibliothekar trifft Sprachtechnologe!

Wurde das nicht schon getan?

- Konkordanzen zwischen kontrollierten Vokabularen
- Mapping von Ontologien
 - Direkt
 - Via Upper-Level Ontologies (SUMO etc.)
- ...

Neu:

- Ziel z.B. für Themen: sehr simple Kategorisierung
 - < 100 Klassen
- Berücksichtigung von existierenden Kategorisierungen in der Sprachtechnologie
- Ziel für Entitätsdefinitionen: wenige, leicht wiederverwendbare Entitätsdefinitionen

Hintergrund: EU-Projekt „META-NET“

- Hauptziel: Langfristige Allianz META für Sprachtechnologie in Europa bauen
 - <http://www.meta-net.eu/join-forms/organisation-form>
- META-NET umfasst mehr als 40 teilnehmende Organisationen aus 30+ Ländern
- META hat bereits über 200 Mitglieder
- Wichtig: Nutzer von und Beiträger zu Sprachtechnologie involvieren – auch Bibliotheken!

META-NET

- Event: META-FORUM 2011
- Budapest, 27.-29. Juni 2011
- Ziel: Nutzer / Sprachtechnologieentwickler / „Entscheider“ zusammenbringen
- Ziele für die Sprachtechnologie in den nächsten 10 Jahren diskutieren
- Details und bald Registrierung unter <http://www.meta-net.eu/events>

Übersicht

- Was ist das mehrsprachige Web?
- Nutzen durch Bibliotheken
- Beiträge durch Bibliotheken

Vielen Dank für Ihre Aufmerksamkeit

