

*Maribor*

# **Big Data in der Translation - Praktische und theoretische Anmerkungen**

*Peter Sandrini*

15/05/2014



# Vortragender

**Position:** Wissenschaftlicher Mitarbeiter  
am Institut für Translationswissenschaft

**Fachgebiete:** Übersetzen von Rechtstexten, Terminologie,  
Translationstechnologie, Globalisierung – Lokalisierung

**Website:** <http://www.petersandrini.net>  
<http://uibk.academia.edu/PeterSandrini>





# Inhaltsübersicht

- 1) Translation 2.0 – Veränderungen
- 2) Big Data: Translationsdaten
- 3) Aussagekraft von Translation Data
- 4) Metadaten für Translation
- 5) Folgen und Auswirkungen

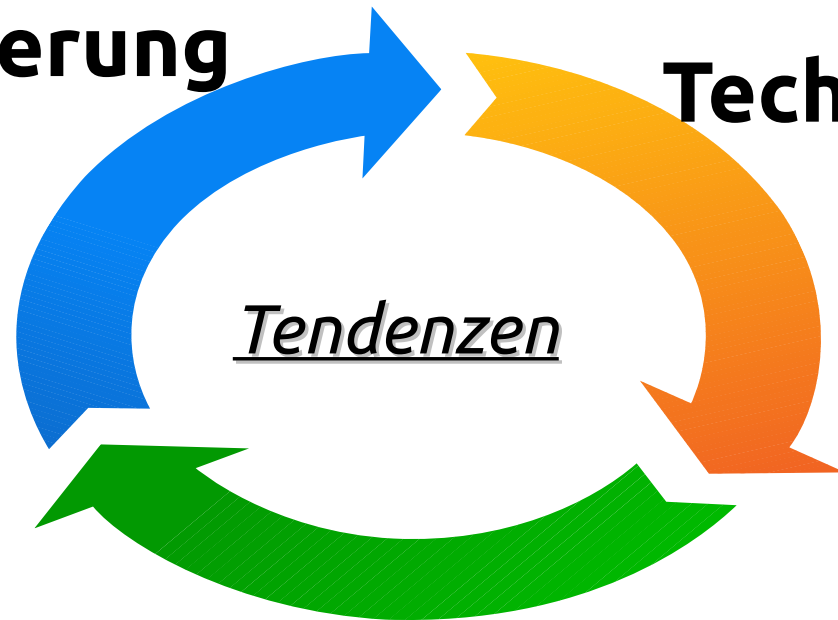


# Translation

# 2.0

**Digitalisierung**

**Technologisierung**



**Globalisierung**

# Digitalisierung

- **digitale Textwelten**

Kopierbarkeit und Transport  
Suchbarkeit und Wiederverwendung

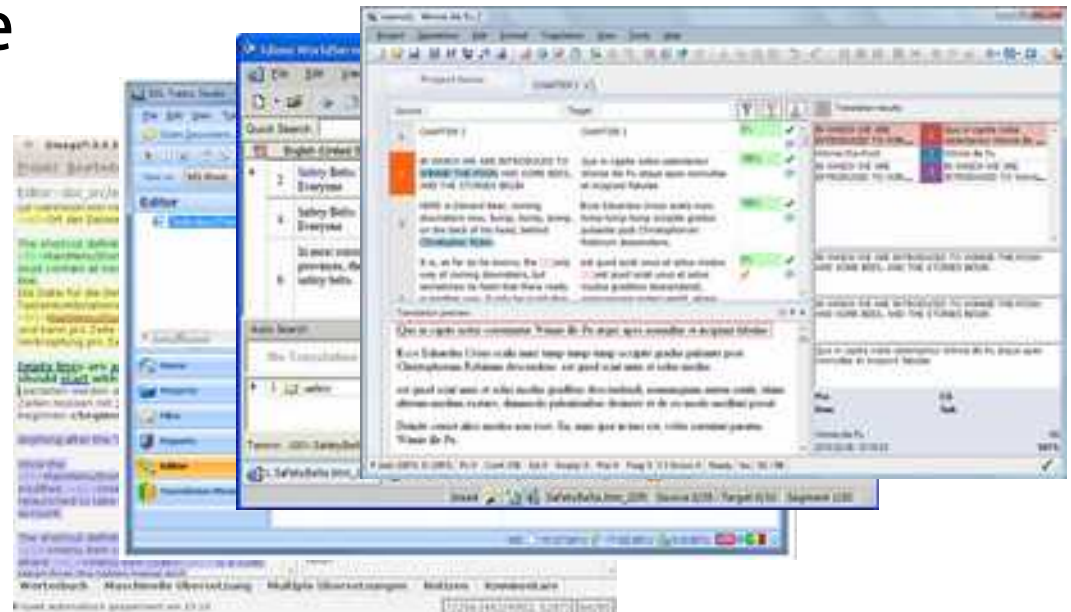
- **digital humanities**

Anwendung von computergestützten Verfahren und  
systematische Verwendung von digitalen Ressourcen  
in den Geistes- und Kulturwissenschaften



# Technologisierung

- Computergestütztes Übersetzen (CAT)
- Einsatz von ‚Translation Environment Tools‘ (TenT)
- Institutionelles und professionelles Übersetzen definieren sich auch über den Einsatz von Translationstechnologie



# Globalisierung und Translation

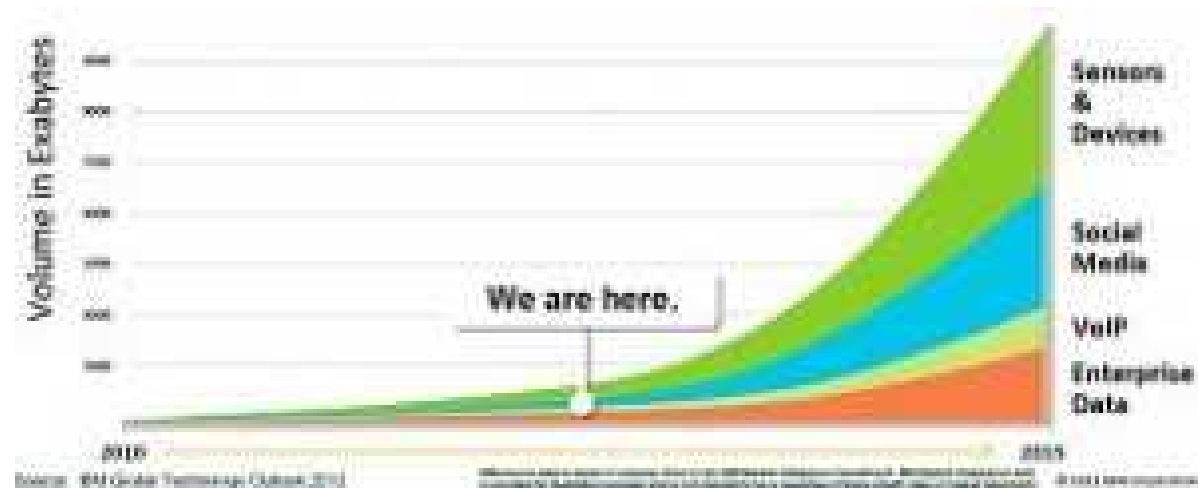
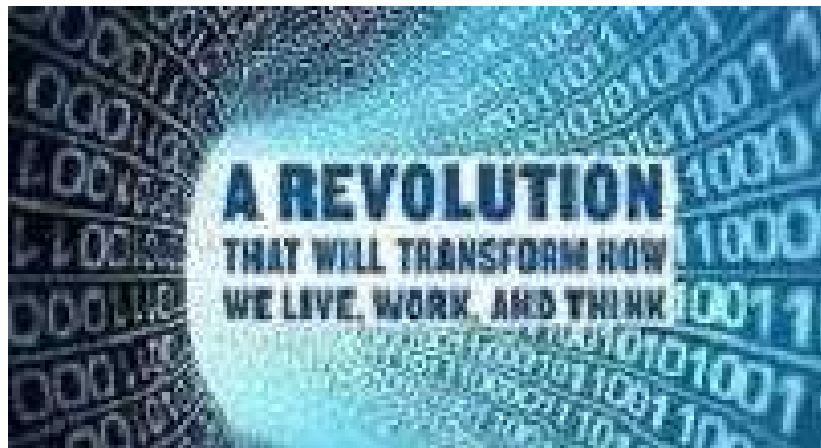
- exponentielles Wachstum des Marktes
- weltweites Kooperationspotenzial
- umfassende Translationsprojekte
- Anforderungen:  
Qualität, zeitgerechte Erledigung,  
Professionalität und  
Wettbewerbsfähigkeit





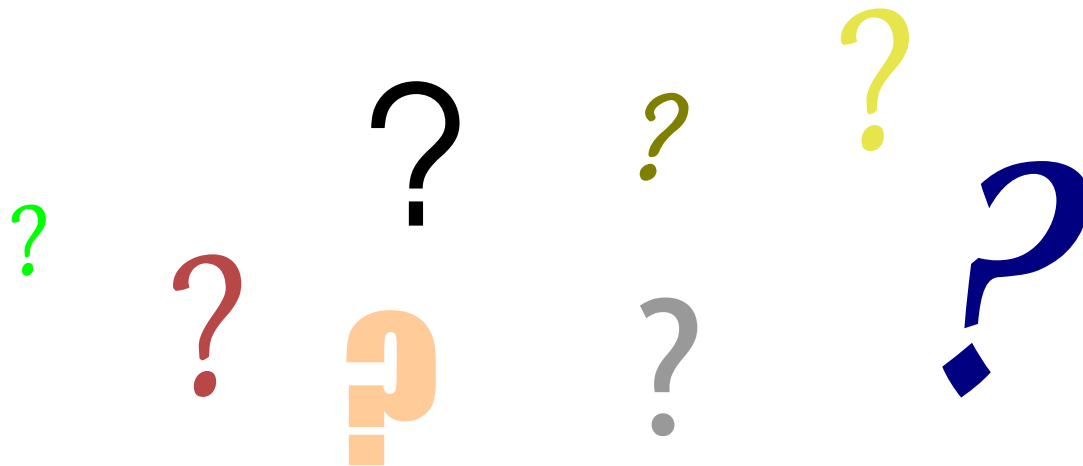
# „datafication“

- “taking information about all things under the sun...and transforming it into a data format to make it quantified”
- datafication: "the idea that more and more businesses are dependent on their data for their business"



# Big Data und Translation

- Gibt es “Datafication” auch in der Translation?
- Von welcher Art von Daten sprechen wir hier?
- Ist Translation zunehmend abhängig von Big Data?
- Was bedeutet das aus praktischer/beruflicher und theoretischer Sicht?



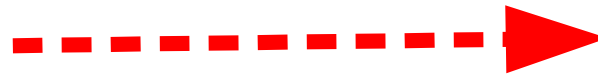
# Translation



- einfache Definition:  
zweckgerichtete Umwandlung eines Ausgangstextes (AT) in einer Sprache in einen Zieltext (ZT) in einer anderen Sprache
- erweiterte Definition:  
"jede konventionalisierte, interlinguale und transkulturelle Interaktion [...], die in einer Kultur als zulässig erachtet wird"  
(Prunc 1997: 108)

# 'datafication' der Translation

*Input*



*Output*



AT

+



Daten

**Translations-  
prozess**

ZT



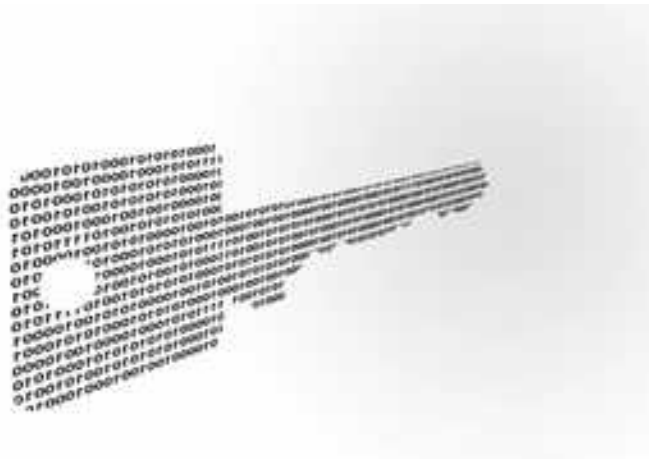
+



Daten

# Daten

- unstrukturierte Daten: z.B. Textmengen, Dokumente, Grafiken
- semistrukturierte Daten: z. B. XML-Daten
- strukturierte Daten: Datenbanken, mit gleichartiger Struktur und Datenmodell



# Translation Data: Definition

- Daten, in denen Übersetzungen in strukturierter und maschinenlesbarer Form gespeichert werden
- unterschiedlich definiert innerhalb der „Translation Process Research“ (TPR) als „Translation Process Data“ bzw. empirische Daten über das Verhalten von Übersetzern während der Übersetzung



# Translation Data: Arten

- Textuelle oder sprachliche Ebenen:
  - Wortebene ➤ lexikographische und terminologische Daten
  - Satzebene ➤ Übersetzungsspeicher (Translation Memory)
  - Textebene ➤ Parallelkorpus







# Translation Data

Die im dem letzten Absatz genannten Unternehmen dürfen als Mitglieder der Genossenschaft aufgenommen werden, vorausgesetzt, dass gegen sie keine Zwangsvollstreckung von beweglichem oder unbeweglichem Vermögen, Proteste, Präventivverleiche oder Konkursverfahren vorliegen, oder dass sie insolvent sind.

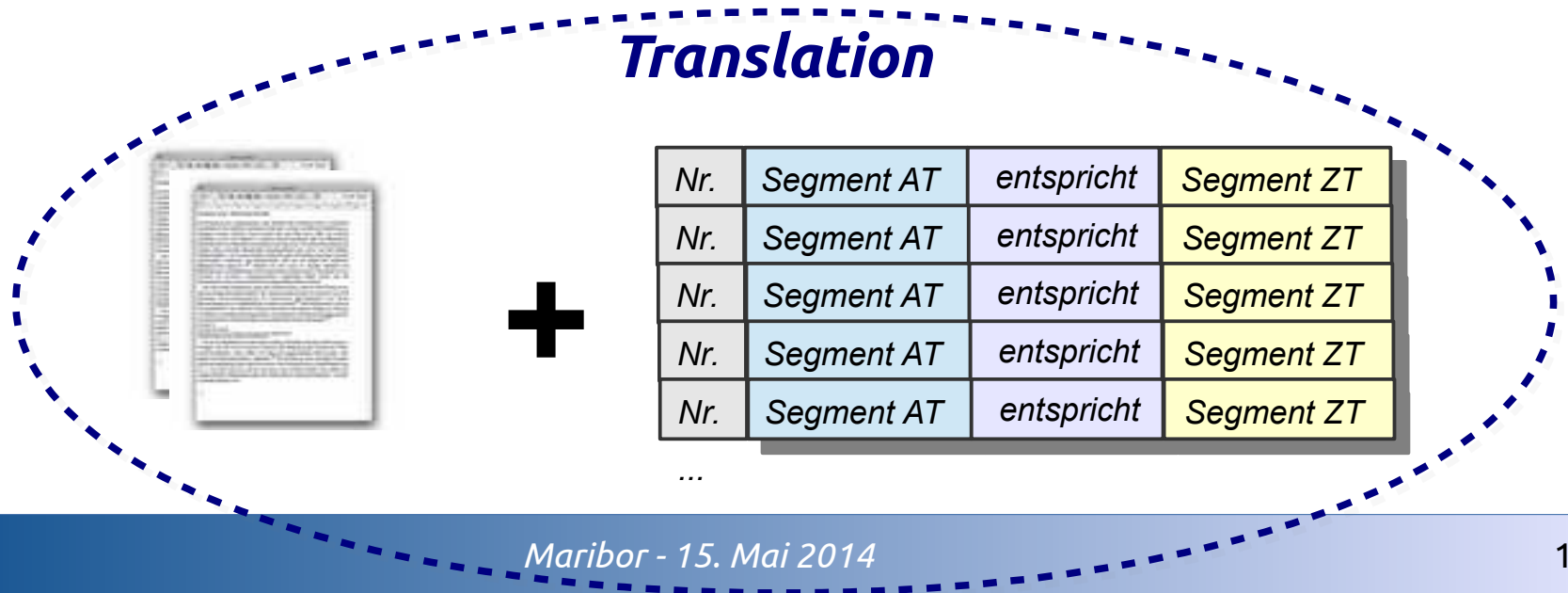
Le imprese di cui al precedente comma possono essere ammesse a socio della Cooperativa purché non abbiano in corso esecuzioni immobiliari e mobiliari e protesti, procedure per concordato preventivo o fallimento, né siano fallite.

=

**Translation Unit**

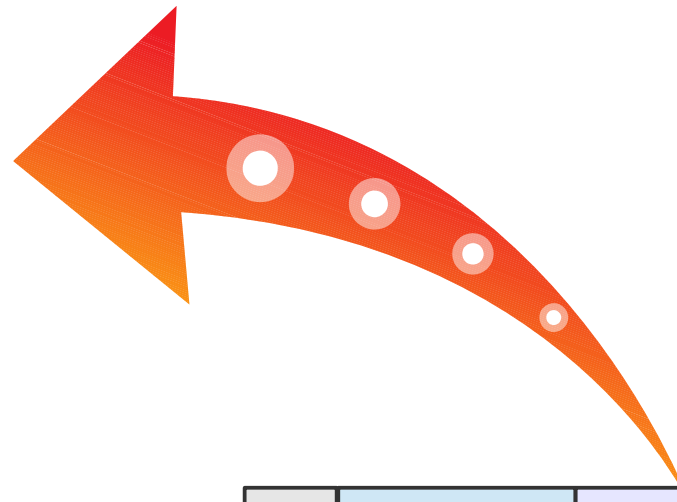
# 'datafication' der Translation

- **Translation** = zweckgerichtetes, sprachliches Übertragen eines Ausgangstextes in einen Zieltext
- **Translation** = Konsultation von Daten + Produktion eines Zieltextes + Produktion von Daten
- **Daten** = (Segmente des) Ausgangstextes + (Segmente des) Zieltextes + Äquivalenzrelationen



# Äquivalenzrelation?

- im Translation-Memory:  
„**das AT-Segment wurde bereits einmal in dieser Form übersetzt**“
- Art der Äquivalenz?
- ...



Nr.	Segment AT	entspricht	Segment ZT
Nr.	Segment AT	entspricht	Segment ZT
Nr.	Segment AT	entspricht	Segment ZT
Nr.	Segment AT	entspricht	Segment ZT
Nr.	Segment AT	entspricht	Segment ZT

# Äquivalenz: Bezugsrahmen nach Koller (1979)

- außersprachlicher Sachverhalt
  - Art der Verbalisierung
  - Text- und Sprachnormen
  - Empfänger/Leser
  - Sprachästhetik
- denotative Äquivalenz
  - konnotative Äquivalenz
  - textnormative Äquivalenz
  - pragmatische Äquivalenz
  - formale Äquivalenz



# Äquivalenz: Bezugsrahmen

- außersprachlicher Sachverhalt
  - Art der Verbalisierung
  - Text- und Sprachnormen
  - Empfänger/Leser
  - Sprachästhetik
  - Ausgangstextfunktion
  - Übersetzungsfunktion
  - ...
- denotative Äquivalenz
  - konnotative Äquivalenz
  - textnormative Äquivalenz
  - pragmatische Äquivalenz
  - formale Äquivalenz
  - textintendierte Äquivalenz
  - finalistische Äquivalenz



*(Königs 1981)*

# Äquivalenz: Objekt

- Relation zwischen Benennungen/Wörtern/Termini
- Relation zwischen Begriffen
- Relation zwischen Teilen verschiedener Sprachsysteme
- Relation zwischen Ausgangstext und Zieltext
- Relation zwischen Texten in Funktion/Situation/Handlung



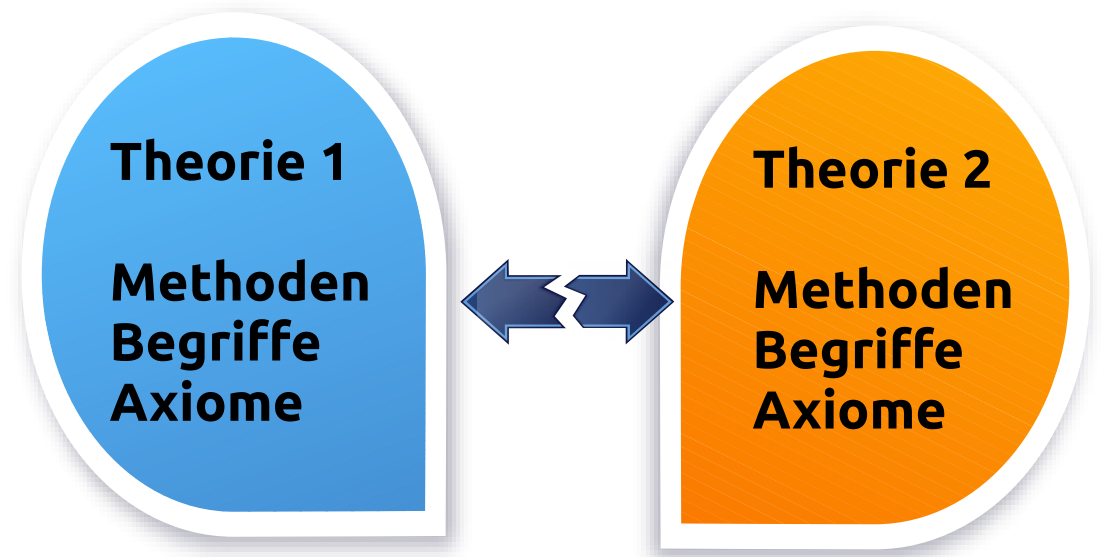
# Äquivalenz: Terminologie

- Bezeichnung eines Begriffs durch zwei Benennungen verschiedener Sprachen  
= Identität der Begriffsinhalte (onomasiologisch)  
≠ Identität der sprachlichen und textuellen Verwendung (semasiologisch)
- Ziel der komparativen Terminologearbeit ist die Identifikation solcher Benennungspaare, die einen einzigen Begriff bezeichnen (*Cole in Lauren/Picht 1993: 400*)

TERM  ology

# Äquivalenz

- „equivalence in difference“ (Jakobson 1959)
- „interchangeable in a given situation“ (Catford 1965)
- „Invarianz auf Inhaltsseite“ (Kade 1968)
- „Wirkungsgleichheit“ (Koller 1972)
- ...





# Adäquatheit

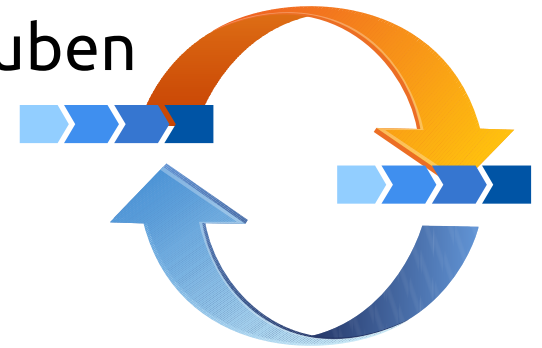
(Reiß/Vermeer 1984)

- Relation zwischen Zieltext und Ausgangstext bei konsequenter Beachtung eines Zweckes (**Skopos**)
- prozessorientiert (i.G. zu produktorientierter Äquivalenz)
- Äquivalenz ist Adäquatheit bei Funktionskonstanz zwischen AT und ZT



# Information in Translation Data

- „das AT-Segment wurde bereits einmal in dieser Form übersetzt“, d.h. das ZT-Segment war damals adäquat
- entspricht Äquivalenzauffassung als historischen Begriff in Toury (2009: 102): „Äquivalenz bezeichnet somit nicht mehr eine einzige Beziehung mit immer derselben Art Invarianz, sondern vielmehr alle Beziehungen, die die Translation unter bestimmten festgelegten Bedingungen charakterisieren“
- Unter welchen Bedingungen? In welchem Kontext? Warum? Für welchen Auftraggeber? In welchem Text/Textsorte? Wann? Für welches Zielpublikum? In welchem Fachgebiet? ...
- zusätzliche Informationen (**Metadaten**) erlauben Rückschlüsse





# Translation Meta-Data: Beispiele

- (Mehr-)Wortebene

*Schlitzsperre*

*briglia a fessura*

**Fachgebiet:**  
**Wildbachverbauung**  
**Geo.: Südtirol**

*Fuß*

*foot*

**Fachgebiet: Anatomie**

*Fuß*

*base*

**Fachgebiet: Technik**

*Fuß*

*root*

**Fachgebiet: Werkzeug**

# Translation Meta-Data: Beispiele

- Satzebene

***Eliminate gender disparity in primary and secondary education, preferably by 2005, and in all levels of education no later than 2015.***

*Das Geschlechtergefälle in der Grund- und Sekundarschulbildung beseitigen, vorzugsweise bis 2005 und auf allen Bildungsebenen bis spätestens 2015.*

**Quelle: UN Millennium Development Goals**

***Eliminate gender disparity in primary and secondary education, preferably by 2005, and in all levels of education no later than 2015.***

*Abbauen der Unterschiede zwischen den Geschlechtern in Primär- und Sekundärschulstufe am besten bis 2005, und auf allen Ebenen der Ausbildung innerhalb 2015.*

# Translation Meta-Data: Beispiele

- Textebene

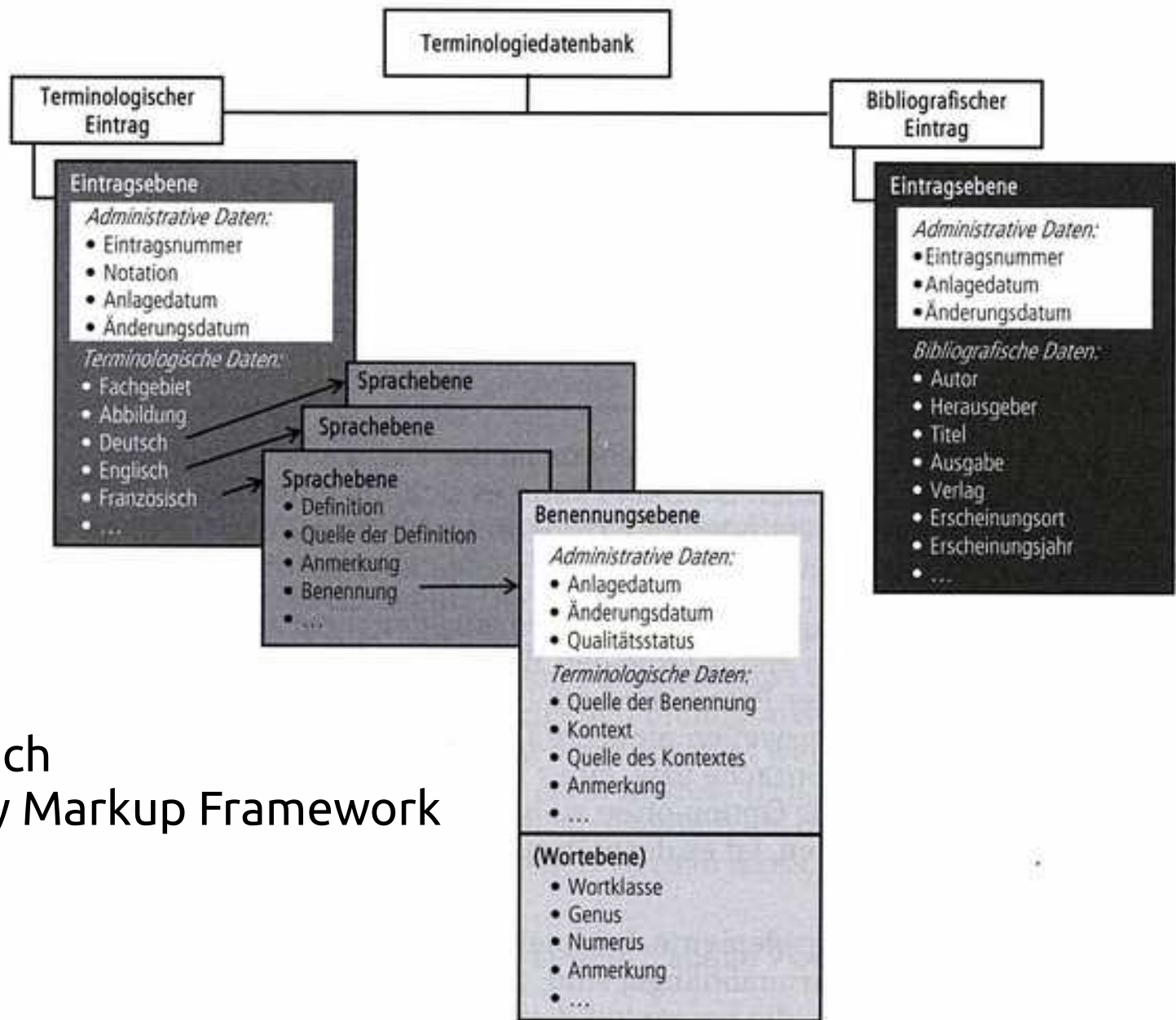
fr	en
<p><b>DIRECTIVE 95/46/CE DU PARLEMENT EUROPÉEN ET DU CONSEIL</b> du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données</p> <p><b>LE PARLEMENT EUROPÉEN ET LE CONSEIL DE L'UNION EUROPÉENNE,</b> vu le traité instituant la Communauté européenne, et notamment son article 100 A, vu la proposition de la Commission (1), vu l'avis du Comité économique et social (2),</p> <p>statuant conformément à la procédure visée à l'article 189 B du traité (3),</p> <p>1.1 Considérant que les objectifs de la Communauté, énoncés dans le traité, ne peuvent être atteints par le traité sur l'union européenne, considéré à l'état actuel, et qu'il est nécessaire d'agir dans ce domaine, notamment à l'égard des relations avec des États tiers, que la Communauté décide, à assurer par une action commune le progrès économique et social en poursuivant les tâches qui incombent à l'Europe, à promouvoir l'harmonisation constante des législations des uns des autres, à préserver et conforter le droit et la liberté, et à promouvoir la démocratie en se fondant sur les droits fondamentaux reconnus dans</p>	<p><b>DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL</b> of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data</p> <p><b>THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,</b> Having regard to the Treaty establishing the European Community, and in particular Article 100a thereof, Having regard to the proposal from the Commission (1), Having regard to the opinion of the Economic and Social Committee (2),</p> <p>Acting in accordance with the procedure referred to in Article 189B of the Treaty (3),</p> <p>1.1 Whereas the objectives of the Community, as laid down in the Treaty, as amended by the Treaty on European Union, include creating an ever closer union among the peoples of Europe, fostering closer relations between the States belonging to the Community, ensuring economic and social progress by common action to eliminate the barriers which divide Europe, encouraging the constant improvement of the living conditions of its peoples, preserving and strengthening peace and liberty and promoting democracy on the basis of the fundamental rights recognized in the constitution and laws of the Member States and</p>

## EU Datenschutzrichtlinie 1995/46/EC

# Translation Meta-Data

- ITS Internationalization Tag Set der W3 Multilingual Web Working Group
- TBX mit TMF
- TMX Tag Set
- XLIFF
- PO comments





Eintragsmodell nach  
TMF - Terminology Markup Framework



# TMX-Metainformationen

- für jede Datei im Header pro TM-Datei

- für jede Übersetzungseinheit (TU)

```
<tu>
```

```
<prop type="x-domain">Medizin</prop>
```

```
<prop type="x-project">Übersetzung02</prop>
```

```
<prop type="x-client">Machines LTD</prop>
```

```
<prop type="x-product">cutter_xy</prop>
```

```
...
```

# TMX-Metainformationen

```
<tu tuid="55_101:6">  
  <prop type="session">55</prop>  
  <prop type="committee">3</prop>  
  <tuv xml:lang="EN">  
    <seg>RESOLUTION 55/101</seg>  
  </tuv>  
  <tuv xml:lang="AR">  
    <seg>55/101 القرار</seg>  
  </tuv>  
  <tuv xml:lang="FR">  
    <seg>RÉSOLUTION 55/101</seg>  
  </tuv>  
  <tuv xml:lang="RU">  
    <seg>РЕЗОЛЮЦИЯ 55/101</seg>  
  </tuv>  
  <tuv xml:lang="ES">  
    <seg>RESOLUCIÓN 55/101</seg>  
  </tuv>  
</tu>
```

```
<tu tuid="55_101:25">  
  <prop type="lead">>true</prop>  
  <tuv xml:lang="EN">  
    <seg>3. Reaffirms that the United Nations shall  
    promote universal respect for and observance of,  
    human rights and fundamental freedoms for all  
    without distinction as to race, sex, language or  
    religion;</seg>  
  </tuv>  
  <tuv xml:lang="AR">  
    <seg>3 - تؤكد من جديد أن الأمم المتحدة ستعمل على أن  
    يشيع في العالم احترام حقوق الإنسان والحريات الأساسية للجميع بلا  
    تمييز بسبب الجنس أو اللغة أو الدين ولا تفريق بين الرجال والنساء،  
    </seg>ومراعاة تلك الحقوق والحريات؛  
  </tuv>  
</tu>
```

[www.uncorpora.org](http://www.uncorpora.org)

# XLIFF - Metadata

- XML Localization Interchange File Format  
Metainformationen möglich jeweils zu

- **<trans-unit>**

Required attributes: id.

Optional attributes: approved, translate, reformat, xml:space , datatype, ts, phase-name , restype, resname, extradata , help-id, menu, menu-option , menu-name, coord, font, css-style, style, exstyle, extype, maxbytes, minbytes , size-unit, maxheight, minheight, maxwidth, minwidth , charclass, non-XLIFF attributes

- **<source>**

xml:lang, ts, non-XLIFF attributes

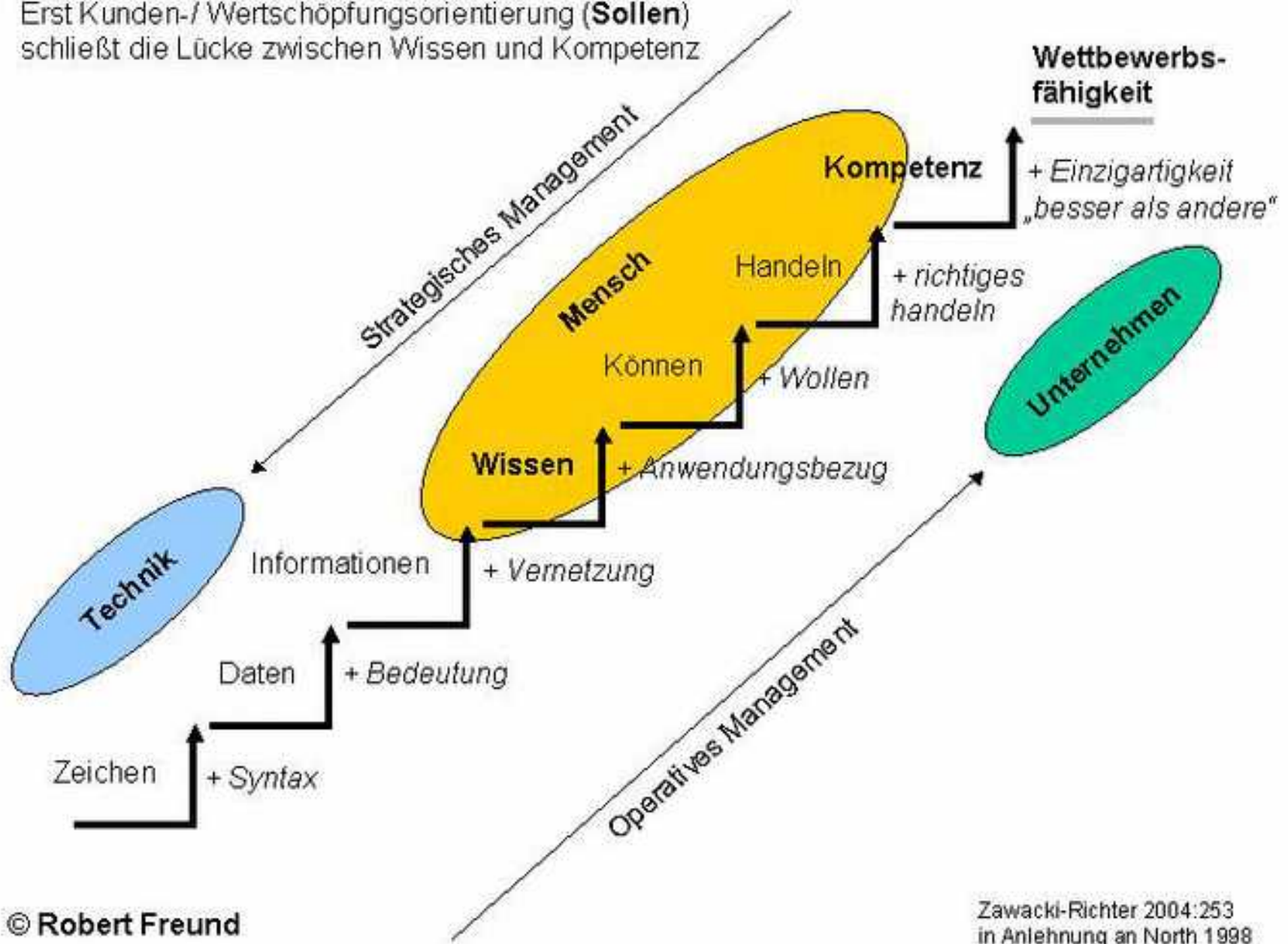
- **<target>**

state, state-qualifier, phase-name, xml:lang, ts, restype, resname, coord, font, css-style, style, exstyle, equiv-trans, non-XLIFF attributes





Erst Kunden-/ Wertschöpfungsorientierung (**Sollen**)  
schließt die Lücke zwischen Wissen und Kompetenz

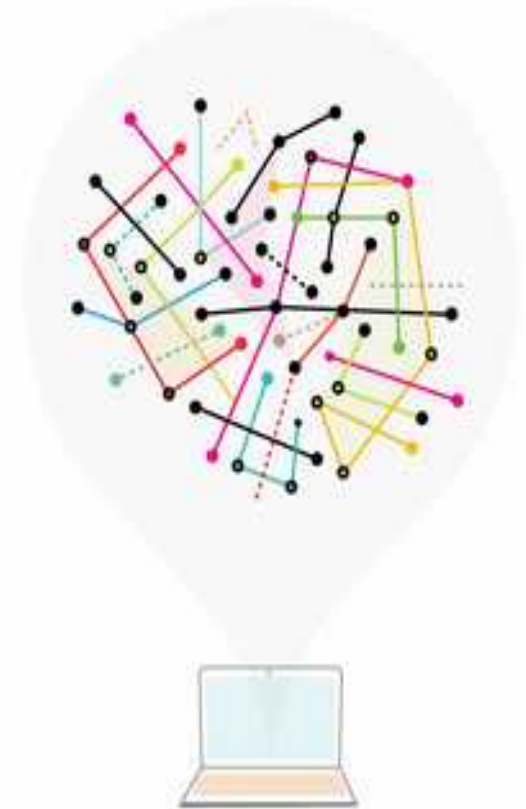


© Robert Freund

Zawacki-Richter 2004:253  
in Anlehnung an North 1998

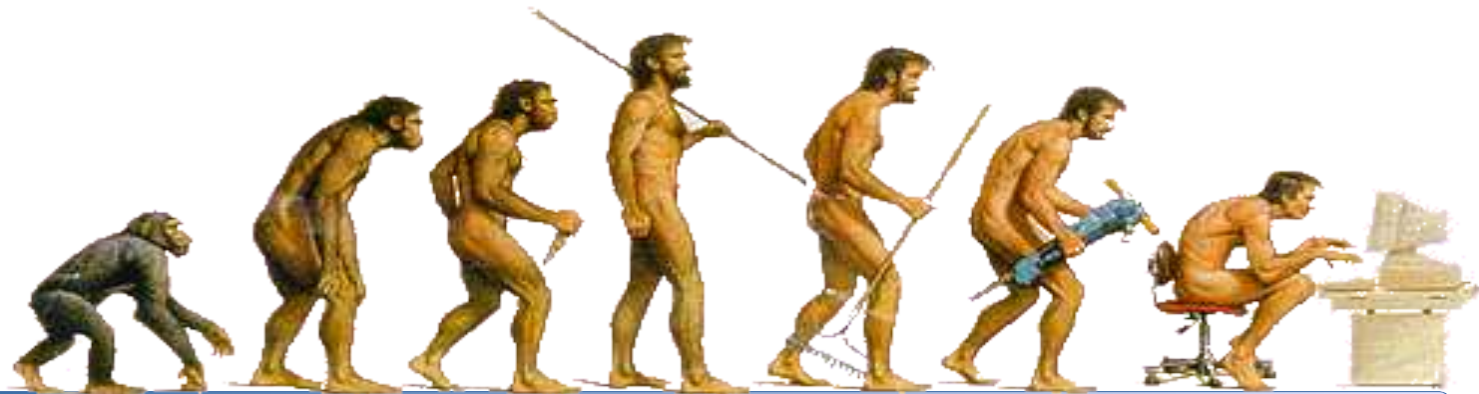
# „Datafication“ of Translation

- **Translation =**
  - Remix (kreatives Neugestalten) aus
  - Daten (Übersetzungsdaten) und
  - Input (Auftragsspezifikation, Skopos)
- **Translation =**
  - Variation des AT +
  - Selektion, Rekombination, Adaptation von Daten



# Translation Data: Folgen

- gewisses Mass an Entpersonifizierung
- Wiederverwenden von Übersetzungen, und dadurch Effizienzsteigerung durch Translation-Memory
- Statistische Maschinenübersetzung (SMT)  
Google translate, Microsoft Translator ...
- ...



# Neue Auffassung von Translation

**Translation** = fallspezifisches Anwenden  
und Produzieren von Translationsdaten



Qualität von Translation = **f** (Translation Data)

aber

Translation Data  $\neq$  automatisches Übersetzen (MT)





# Neue Auffassung von Sprache

## Was ist Sprache:

Eine repräsentative Sammlung von sprachlichen Daten,  
kein begrenztes, regelbasiertes System



- Empirische Sprachverwendungsforschung
- Sprachkorpora
- Statistische Maschinenübersetzung
- ...



# Translation Data und Sprachkompetenz?

- Sprach- und Translationskompetenz notwendig zur Beurteilung, ob TD eingesetzt werden können
- Sprach- und Translationskompetenz notwendig zur zielgerechten Verwendung von TD
- Sprach- und Translationskompetenz notwendig zur fallspezifischen Evaluierung von MT-Output
- ...





## **Website:**

<http://www.petersandrini.net>

<http://uibk.academia.edu/PeterSandrini>