



## Netzbasierte Informationssysteme **Mehrsprachigkeit im Web**

Prof. Dr.-Ing. Robert Tolksdorf  
Freie Universität Berlin  
Institut für Informatik  
Netzbasierte Informationssysteme  
mailto: [tolk@inf.fu-berlin.de](mailto:tolk@inf.fu-berlin.de)  
<http://www.robert-tolksdorf.de>

- Sprachen im Web
- Bezeichnung von Sprachen
- Markierung sprachlicher Eigenschaften
- Zeicheneigenschaften



## Mehrsprachigkeit im Web Zeichen, Schriften, Sprachen

# Mehrsprachige Seiten

- Unterschiede in
  - Sprache
  - Schriftzeichen
  - Schriftcodierung
  - Schreibrichtung
  - Kulturellen Konventionen
  - ...



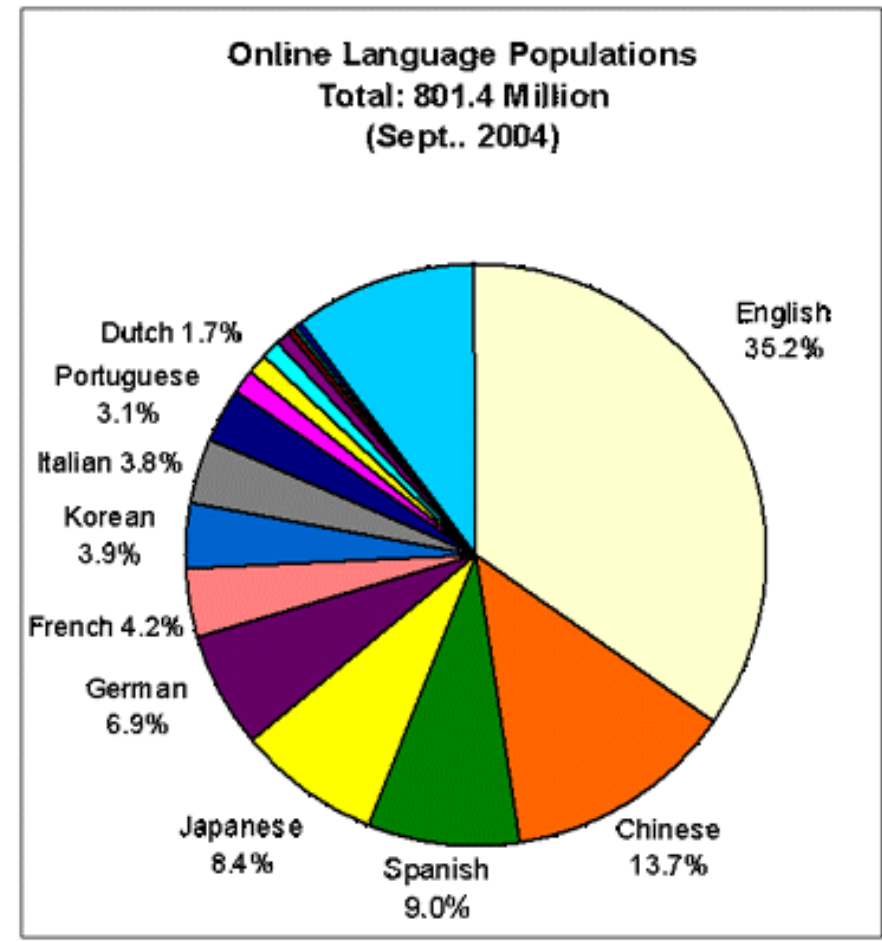
# Sprachen der Welt (2005)

Language	Native	2nd	Total
Mandarin Chinese	873 million	178 million	1051 million
Hindi	370 million	120 million	490 million
English	340 million		510 million
Spanish	350 million	70 million	420 million
Arabic	206 million	24 million	230 million
Portuguese	203 million	10 million	213 million
Bengali	196 million		215 million
Russian	145 million	110 million	255 million
Japanese	126 million	1 million	127 million
German	101 million	128 million	129 million

[Vistawide. [http://www.vistawide.com/languages/top\\_30\\_languages.htm](http://www.vistawide.com/languages/top_30_languages.htm)]

# Sprachen der Online-Population

	Internet access (M)	%'age world online population
English	287.5	35.8%
<i>Non-English</i>	516.7	64.2%
TOTAL EUROPEAN LANGUAGES (excl. English)	276.0	37.9%
TOTAL ASIAN LANGUAGES	240.6	33.0%
TOTAL WORLD	729.2	



[Quelldaten nicht konsistent!]

[Quelle: Global Reach (global-reach.biz/globstats), 9/04]

# Internationalisierung

- *Internationalisierung* ist die Planung und Implementierung von Diensten und Produkten so dass sie einfach an lokale Sprachen und Kulturen anpassbar sind, was *Lokalisierung* ist
- Internationalisierung
  - „I18N“ - „I - eighteen letters –N“ – „Internationalization“
  - Voraussetzung für Lokalisierung
  - Beispiele
    - Platzgestaltung in GUIs läßt Raum für Sprachen die mehr Zeichen benötigen
    - Verwendung internationaler Zeichenrepertoires und -codes, z.B. Unicode
    - Vergabe leicht übersetzbarer Beschreibungen für Graphiken
    - Verwendung allgemeinverständlicher Beispiele (Social Security Number ...)
    - Vorausplanung der Übersetzung in Sprachen mit Kodierungen mit mehr als einem Byte pro Zeichen in Software

# Lokalisierung

---

- *Lokalisierung* ist die Anpassung eines Produktes oder Dienstes an eine Sprache, Kultur und lokales "look-and-feel" was durch *Internationalisierung* vereinfacht wird
- Lokalisierung
  - „L10N“ – „L - ten letters –N“ – „Localization“
  - Übersetzung
  - Aber auch: Anpassung an Zeitzonen, Währung, Feiertage, Farbkonventionen, Namen, Geschlechterrollen etc.
  - Ziel: Lokalisiertes Produkt oder Dienst soll so aussehen, als sei er/es lokal entwickelt worden





## Bezeichnung von Sprachen

# Sprachbezeichner

---

- Sprachen im Internet durch Codes bezeichnet
- Basis nach RFC 3066 (früher 1766)
  - In ISO 639 definierte Kürzel für Sprachen
  - In ISO 3166 definierte Kürzel für Länder
- Format
  - Sprachcode: de en etc.
  - Sprachcode-Ländercode: de-ch en-uk
    - Matching nach Substring am Anfang en passt auf en-us
    - Groß-/Kleinschreibung irrelevant en passt auf En-us und EN
  - Experimentell: x-kl i ngon  
(siehe auch <http://www.google.com/intl/xx-klingon/>)
  - IANA basiert: i-klingon
- Nicht perfekt: Lateinamerikanisches Spanisch?

# Sprachcodes nach ISO 639

aa	Afar	eu	Baskisch	kl	Grönländisch	or	Orija	ta	Tamilisch
ab	Abchasisch	fa	Persisch	km	Kambodschanisch	pa	Pundjabisch	te	Telugu
af	Afrikaans	fi	Finnisch	kn	Kannada	pl	Polnisch	tg	Tadschikisch
am	Amharisch	fj	Fiji	ko	Koreanisch	ps	Paschtu	th	Thai
ar	Arabisch	fo	Faröisch	ks	Kaschmirisch	pt	Portugiesisch	ti	Tigrinja
as	Assamesisch	fr	Französisch	ku	Kurdisch	qu	Quechua	tk	Turkmenisch
ay	Aymara	fy	Friesisch	ky	Kirgisisch	rm	Rätoromanisch	tl	Tagalog
az	Aserbaidtschanisch	ga	Irish	la	Lateinisch	rn	Kirundisch	tn	Sezuan
ba	Baschkirisch	gd	Schottisches Gälisch	ln	Lingalisch	ro	Rumänisch	to	Tongaisch
be	Belorussisch	gl	Galizisch	lo	Laotisch	ru	Russisch	tr	Türkisch
bg	Bulgarisch	gn	Guarani	lt	Litauisch	rw	Kijarwanda	ts	Tsongaisch
bh	Biharisch	gu	Gujaratisch	lv	Lettisch	sa	Sanskrit	tt	Tatarisch
bi	Bislamisch	ha	Hausa	mg	Malagasisch	sd	Zinti	tw	Twi
bn	Bengalisch	he (iw)	Hebräisch	mi	Maorisch	sg	Sango	uk	Ukrainisch
bo	Tibetanisch	hi	Hindi	mk	Mazedonisch	sh	Serbokroatisch	ur	Urdu
br	Bretonisch	hr	Kroatisch	ml	Malajalam	si	Singhalesisch	uz	Usbekisch
ca	Katalanisch	hu	Ungarisch	mn	Mongolisch	sk	Slowakisch	vi	Vietnamesisch
co	Korsisch	hy	Armenisch	mo	Moldavisch	sl	Slowenisch	vo	Volapük
cs	Tschechisch	ia	Interlingua	mr	Marathi	sm	Samoanisch	wo	Wolof
cy	Walisisch	id (in)	Indonesisch	ms	Malaysisch	sn	Schonisch	xh	Xhosa
da	Dänisch	ie	Interlingue	mt	Maltesisch	so	Somalisch	yi (ji)	Jiddish
de	Deutsch	ik	Inupiak	my	Burmesisch	sq	Albanisch	yo	Joruba
dz	Bhutani	is	Isländisch	na	Nauruisch	sr	Serbisch	zh	Chinesisch
el	Griechisch	it	Italienisch	ne	Nepalisch	ss	Swasiländisch	zu	Zulu
en	Englisch	ja	Japanisch	nl	Holländisch	st	Sesothisch		
eo	Esperanto	jw	Javanisch	no	Norwegisch	su	Sudanesisch		
es	Spanisch	ka	Georgisch	oc	Okzitanisch	sv	Schwedisch		
et	Estnisch	kk	Kasachisch	om	Oromo	sw	Suaheli		

# Ländercodes nach ISO 3166

---

AFGHANISTAN	AF	BRITISH INDIAN OCEAN TERRITORY	IO
ALBANIA	AL	BRUNEI DARUSSALAM	BN
ALGERIA	DZ	BULGARIA	BG
AMERICAN SAMOA	AS	BURKINA FASO	BF
ANDORRA	AD	BURUNDI	BI
ANGOLA	AO	CAMBODIA	KH
ANGUILLA	AI	CAMEROON	CM
ANTARCTICA	AQ	CANADA	CA
ANTIGUA AND BARBUDA	AG	CAPE VERDE	CV
ARGENTINA	AR	CAYMAN ISLANDS	KY
ARMENIA	AM	CENTRAL AFRICAN REPUBLIC	CF
ARUBA	AW	CHAD	TD
AUSTRALIA	AU	CHILE	CL
AUSTRIA	AT	CHINA	CN
AZERBAIJAN	AZ	CHRISTMAS ISLAND	CX
BAHAMAS	BS	COCOS (KEELING) ISLANDS	CC
BAHRAIN	BH	COLOMBIA	CO
BANGLADESH	BD	COMOROS	KM
BARBADOS	BB	CONGO	CG
BELARUS	BY	CONGO, THE DEMOCRATIC REPUBLIC OF THE	CD
BELGIUM	BE	COOK ISLANDS	CK
BELIZE	BZ	COSTA RICA	CR
BENIN	BJ	CÔTE D'IVOIRE	CI
BERMUDA	BM	CROATIA	HR
BHUTAN	BT	CUBA	CU
BOLIVIA	BO	CYPRUS	CY
BOSNIA AND HERZEGOVINA	BA	CZECH REPUBLIC	CZ
BOTSWANA	BW	DENMARK	DK
BOUVET ISLAND	BV	DJIBOUTI	DJ
BRAZIL	BR	DOMINICA	DM

---

# Ländercodes nach ISO 3166

DOMINICAN REPUBLIC	DO	GUINEA-BISSAU	GW
EAST TIMOR	TL	GUYANA	GY
ECUADOR	EC	HAITI	HT
EGYPT	EG	HEARD ISLAND AND MCDONALD ISLANDS	HM
EL SALVADOR	SV	HOLY SEE (VATICAN CITY STATE)	VA
EQUATORIAL GUINEA	GQ	HONDURAS	HN
ERITREA	ER	HONG KONG	HK
ESTONIA	EE	HUNGARY	HU
ETHIOPIA	ET	ICELAND	IS
FALKLAND ISLANDS (MALVINAS)	FK	INDIA	IN
FAROE ISLANDS	FO	INDONESIA	ID
FIJI	FJ	IRAN, ISLAMIC REPUBLIC OF	IR
FINLAND	FI	IRAQ	IQ
FRANCE	FR	IRELAND	IE
FRENCH GUIANA	GF	ISRAEL	IL
FRENCH POLYNESIA	PF	ITALY	IT
FRENCH SOUTHERN TERRITORIES	TF	JAMAICA	JM
GABON	GA	JAPAN	JP
GAMBIA	GM	JORDAN	JO
GEORGIA	GE	KAZAKHSTAN	KZ
GERMANY	DE	KENYA	KE
GHANA	GH	KIRIBATI	KI
GIBRALTAR	GI	KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF	KP
GREECE	GR	KOREA, REPUBLIC OF	KR
GREENLAND	GL	KUWAIT	KW
GRENADA	GD	KYRGYZSTAN	KG
GUADELOUPE	GP	LAO PEOPLE'S DEMOCRATIC REPUBLIC	LA
GUAM	GU	LATVIA	LV
GUATEMALA	GT	LEBANON	LB
GUINEA	GN	LESOTHO	LS

# Ländercodes nach ISO 3166

LIBERIA	LR	NETHERLANDS	NL
LIBYAN ARAB JAMAHIRIYA	LY	NETHERLANDS ANTILLES	AN
LIECHTENSTEIN	LI	NEW CALEDONIA	NC
LITHUANIA	LT	NEW ZEALAND	NZ
LUXEMBOURG	LU	NICARAGUA	NI
MACAO	MO	NIGER	NE
MACEDONIA, THE FORMER YUGOSLAV REPUBLIC OF	MK	NIGERIA	NG
MADAGASCAR	MG	NIUE	NU
MALAWI	MW	NORFOLK ISLAND	NF
MALAYSIA	MY	NORTHERN MARIANA ISLANDS	MP
MALDIVES	MV	NORWAY	NO
MALI	ML	OMAN	OM
MALTA	MT	PAKISTAN	PK
MARSHALL ISLANDS	MH	PALAU	PW
MARTINIQUE	MQ	PALESTINIAN TERRITORY, OCCUPIED	PS
MAURITANIA	MR	PANAMA	PA
MAURITIUS	MU	PAPUA NEW GUINEA	PG
MAYOTTE	YT	PARAGUAY	PY
MEXICO	MX	PERU	PE
MICRONESIA, FEDERATED STATES OF	FM	PHILIPPINES	PH
MOLDOVA, REPUBLIC OF	MD	PITCAIRN	PN
MONACO	MC	POLAND	PL
MONGOLIA	MN	PORTUGAL	PT
MONTSERRAT	MS	PUERTO RICO	PR
MOROCCO	MA	QATAR	QA
MOZAMBIQUE	MZ	RÉUNION	RE
MYANMAR	MM	ROMANIA	RO
NAMIBIA	NA	RUSSIAN FEDERATION	RU
NAURU	NR	RWANDA	RW
NEPAL	NP	SAINT HELENA	SH

# Ländercodes nach ISO 3166

SAINT KITTS AND NEVIS	KN	THAILAND	TH
SAINT LUCIA	LC	TOGO	TG
SAINT PIERRE AND MIQUELON	PM	TOKELAU	TK
SAINT VINCENT AND THE GRENADINES	VC	TONGA	TO
SAMOA	WS	TRINIDAD AND TOBAGO	TT
SAN MARINO	SM	TUNISIA	TN
SAO TOME AND PRINCIPE	ST	TURKEY	TR
SAUDI ARABIA	SA	TURKMENISTAN	TM
SENEGAL	SN	TURKS AND CAICOS ISLANDS	TC
SEYCHELLES	SC	TUVALU	TV
SIERRA LEONE	SL	UGANDA	UG
SINGAPORE	SG	UKRAINE	UA
SLOVAKIA	SK	UNITED ARAB EMIRATES	AE
SLOVENIA	SI	UNITED KINGDOM	GB
SOLOMON ISLANDS	SB	UNITED STATES	US
SOMALIA	SO	UNITED STATES MINOR OUTLYING ISLANDS	UM
SOUTH AFRICA	ZA	URUGUAY	UY
SOUTH GEORGIA AND THE SOUTH SANDWICH ISLANDS	GS	UZBEKISTAN	UZ
SPAIN	ES	VANUATU	VU
SRI LANKA	LK	VENEZUELA	VE
SUDAN	SD	VIET NAM	VN
SURINAME	SR	VIRGIN ISLANDS, BRITISH	VG
SVALBARD AND JAN MAYEN	SJ	VIRGIN ISLANDS, U.S.	VI
SWAZILAND	SZ	WALLIS AND FUTUNA	WF
SWEDEN	SE	WESTERN SAHARA	EH
SWITZERLAND	CH	YEMEN	YE
SYRIAN ARAB REPUBLIC	SY	YUGOSLAVIA	YU
TAIWAN, PROVINCE OF CHINA	TW	ZAMBIA	ZM
TAJIKISTAN	TJ	ZIMBABWE	ZW
TANZANIA, UNITED REPUBLIC OF	TZ		

# Sprachkürzel nach RFC 4646

- RFC 3066
  - Sprachkürzel:  
Ländercode-Sprachcode (en-US)
  - Bezüge auf ISO Standards
- RFC 4646
  - Sprachkürzel:  
language-script-region-variant-extension-privateuse
  - <http://www.iana.org/assignments/language-subtag-registry>:
    - Type: language  
Subtag: fr  
Description: French  
Added: 2005-10-16  
Suppress-Script: Latn
    - Type: region  
Subtag: CA  
Description: Canada  
Added: 2005-10-16
  - fr-CA ist gültiges Sprachkürzel



# Sprachkürzel nach RFC 4646

- language Kürzel
  - Sprachkürzel, zwei oder drei Buchstaben
- script Kürzel
  - Schreibschrift
  - az-Latn  
Aserbaidshisch in lateinischer Schrift
- region Kürzel
  - Region
  - es-005: Spanisch in Südamerika
  - zh-Hant-HK: Chinesisch in traditioneller Schreibweise in Hong-Kong
  - Regionen nicht nur Länder (ISO 3166)
  - UNM.49 Codes

# Sprachkürzel nach RFC 4646

---

- Variant Kürzel
  - de-CH-1901: Deutsch nach der Reform von 1901 in der Schweiz...
  - sl-rozaj: Dialekt von Slovenisch
- Extension Kürzel
  - Für spätere Erweiterungen von RFC 4646
- Private-use Kürzel
  - Lokale Kürzel



## Markierung sprachlicher Eigenschaften

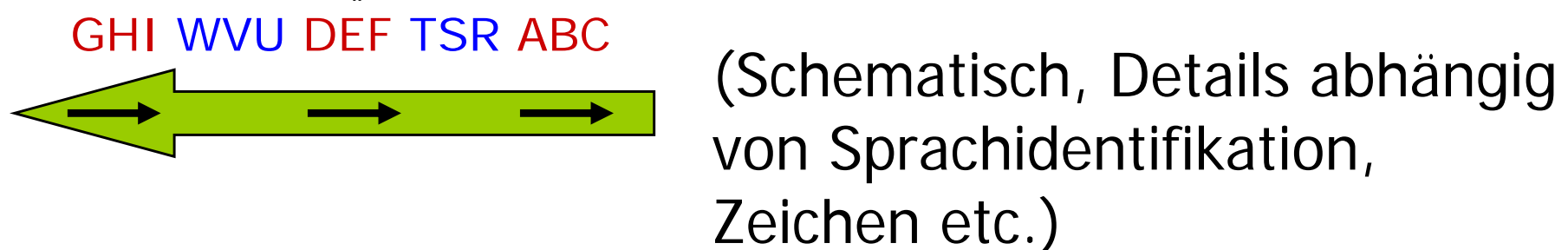
- Alle HTML Elemente können Sprachbezogene Attribute tragen
  - lang-Attribut: Wert ist Sprachcode
    - Wird vom umgebenden Element „geerbt“
    - Kann jeweils überschrieben werden
    - Default ist durch Content-language HTTP Header gegeben
  - dir-Attribut: (Horizontale) Schreibrichtung der Schrift
    - ltr: Left-to-Right
    - rtl: Right-to-Left
    - Wird vom umgebenden Element „geerbt“
    - Kann jeweils überschrieben werden
- Falls nicht an anderes Element gebunden mit `<span>` (oder `<div>`) verwenden

- **ABC, DEF, GHI** aus Schrift, die rechts nach links geschrieben wird (mit `<span dir="ltr">` markiert)
- **RST, UVW** aus Schrift, die links nach rechts geschrieben wird (mit `<span dir="rtl">` markiert)
- `<span dir="ltr">ABC</span> <span dir="rtl">RST</span> <span dir="ltr">DEF</span> <span dir="rtl">UVW</span> GHI`
- Zwei Möglichkeiten, UNICODE Bidirectional Algorithm

- `<html dir="ltr">`:



- `<html dir="rtl">`:



- Was ist die Schreibrichtung des Quelltextes?
- Falls schon visuell geordnet, dann versagt Verarbeitung der Richtungsangaben

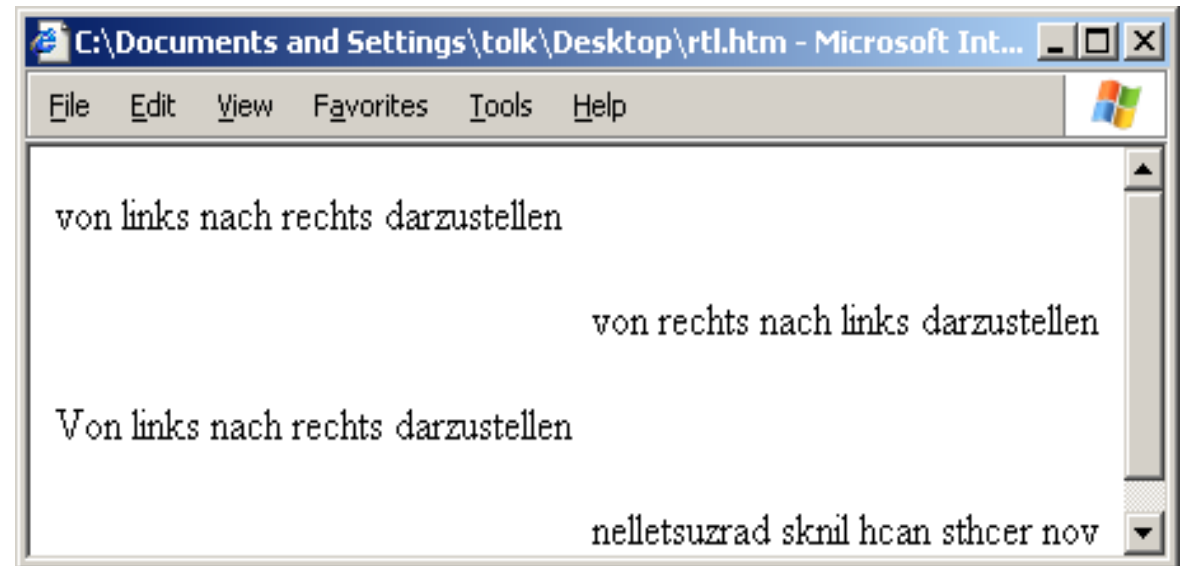
- Bidirectional Algorithm Override, <bdo>-Tag:

```
<p dir="ltr">von links nach rechts darzustellen</p>
```

```
<p dir="rtl">von rechts nach links darzustellen</p>
```

```
<p dir="ltr"><bdo>Von links nach rechts  
darzustellen</bdo></p>
```

```
<p dir="rtl"><bdo>von  
rechts nach links  
darzustellen</bdo>  
</p>
```



(Beispiel ist deutschsprachig,  
daher nur andere Ausrichtung  
bei dir="rtl")

- „Ruby“ ist erklärende Annotation für einen anderen Text

- 

World Wide Web ← *ruby text*  
W W W ← *ruby base*

新幹線 ← *ruby base*  
shinkansen ← *ruby text*

しんかんせん ← *ruby text*  
新幹線 ← *ruby base*

```
<ruby>
```

```
  <rb>WWW</rb>
```

```
  <rt>World Wide Web</rt>
```

```
</ruby>
```

# Spracheigenschaften in CSS

- In CSS2 neue Pseudoklasse :lang
  - :lang(en) {color: red}
  - :lang(fr) {color: blue}
  - Noch nicht implementiert
- In CSS2 Selektorenausdrücke auf Inhalt des lang Attributs
  - \*[lang|=en] {color: red}
  - \*[lang|=fr] {color: blue} Ein Absatz mit einem **chaotic** Sprachgebrauch **ridicule**.

```
<p>Ein Absatz mit einem <span lang="en-us">chaoti c</span>
Sprachgebrauch <span lang="fr">ri di cul e</span>. </p>
```
- Eigenschaft direction mit Werten ltr und rtl
- Eigenschaft unicode-bidi
  - Werte normal, embed, bidi-override
  - <bdo>=unicode-bidi: bidi-override



# CSS2: Anführungszeichen

- Eigenschaft `quotes` legt doppelte und einfache An- und Abführungszeichen fest
- Kombiniert mit `lang` Pseudoelementen:
 

```
Q: lang(en) { quotes: '""' '""' '""' '""' }
Q: lang(no) { quotes: "«" "»" "<" ">" }
```
- Als `open-quote` und `close-quote` verwendbar:
 

```
Q: before { content: open-quote }
Q: after  { content: close-quote }
```
- Sprachabhängige Zitatmarkierung:
 

```
<HTML lang="no" >
  <HEAD>...</HEAD>
  <BODY>
    <P><Q>Trøndere gråter når <Q>Vi nsjan på kai a</Q> bli r
    dekl amert. </Q>
  </BODY>
</HTML>
```

«Trøndere gråter når <Vinsjan på kaia> blir deklamert.»

# Sprachabhängige Anführungszeichen

---

- »Dansk ´da´ Dänisch«
- „Deutsch `de´“
- “English `en´ Englisch”
- « Français « fr » Französisch »
- « Italiano «it» Italienisch»
- «Norsk ´no´ Norwegisch»
- «„ru” Russisch »

# Spracheigenschaften in CSS2

---

- `list-style-type` Eigenschaft legt die Nummerierung von Listen fest
- In CSS1  
`disc`, `circle`, `square`, `decimal`, `lower-roman`,  
`upper-roman`, `lower-alpha`, `upper-alpha`, `none`
- In CSS2 zusätzlich
  - `hebrew` Hebräisch
  - `georgian` Georgisch (`an`, `ban`, `gan`, ..., `he`, `tan`, `in`, `in-an`, ...).
  - `hiragana` `a`, `i`, `u`, `e`, `o`, `ka`, `ki`, ...
  - `katakana` `A`, `I`, `U`, `E`, `O`, `KA`, `KI`, ...
  - `hiragana-iroha` `i`, `ro`, `ha`, `ni`, `ho`, `he`, `to`, ...
  - `katakana-iroha` `I`, `RO`, `HA`, `NI`, `HO`, `HE`, `TO`, ...
  - weitere

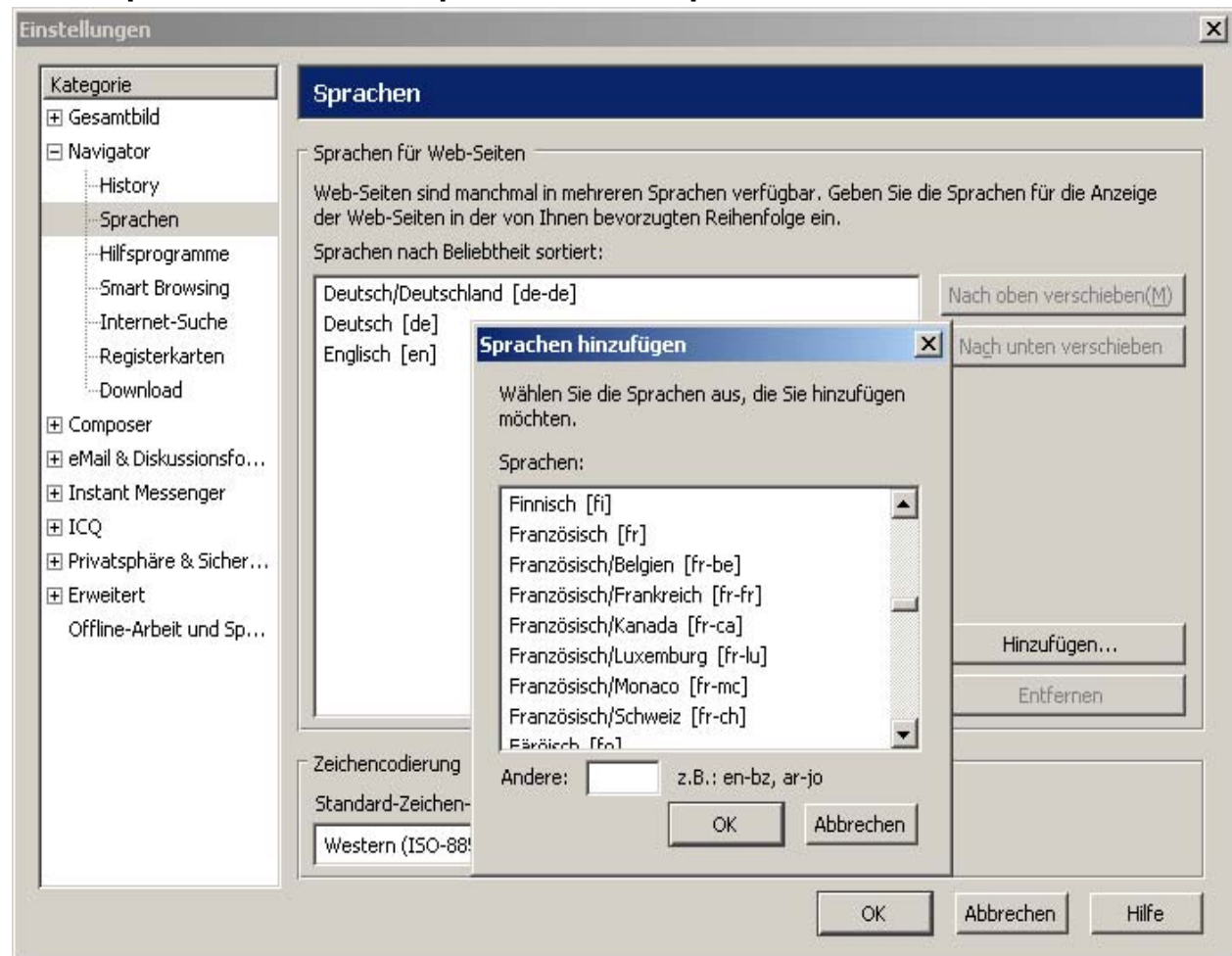
# Spracheigenschaften in XML

---

- Attribut `xml : lang` in XML definiert, also immer verfügbar
- Bedeutung wie `lang` in HTML
- In XHTML sowohl `xml : lang` als auch `lang` benutzen

# Sprache in HTTP

- Browser kann Präferenzen im HTTP-Request mitteilen:  
 GET / HTTP/1.1  
 AcceptLanguage: en-us; q=0.75, en; q=0.5; \*; q=0.25
- q gibt  
 Priorität an,  
 \* ist  
 Platzhalter
- Vom Browser  
 abhängig:



# Sprache in HTTP

---

- Server teilt Encoding in Antwort mit  
200 OK HTTP/1.1  
Content-Language: fr

<html ...

...



## Zeicheneigenschaften

# Zeicheneigenschaften

- Zeichenrepertoire (Character Set, Abstract Character Repertoire, ACS)
  - Eine Menge von Zeichen
  - Definiert durch Namen und Beispiele
  - {Pfund (£), Zett (Z), Ypsilon (Y), Herz (♥)}
  - Keine Ordnung, keine Codierung
- Zeichencode (Coded Character Set, CCS)
  - Abbildung(en) Zeichen → Zeichenposition
  - Z → 5A, ç → FEA5 (Khah)
  - z.B. UNICODE, ISO 8859-1



# UNICODE: Braille

28B3		28C3		288	289	28A	28B	28C	28D	28E	28F	280	281	282	283	284	285	286	287	
				0																
				1																
				2																
				3																
				4																
				5																
				6																
				7																
				8																
				9																
				A																
				B																
				C																
				D																
				E																
				F																

# UNICODE: Beschreibung der Tibetanischen Zeichen

## 0F36

- 0F36 ࠄ TIBETAN MARK CARET -DZUD  
RTAGS BZHI MIG CAN  
• marks point of text insertion or annotation
- 0F37 ࠅ TIBETAN MARK NGAS BZUNG SGOR  
RTAGS  
• emphasis; used like underlining
- 0F38 ࠆ TIBETAN MARK CHE MGO
- 0F39 ࠇ TIBETAN MARK TSA -PHRU  
• a lenition mark

### Paired punctuation

- 0F3A ࠈ TIBETAN MARK GUG RTAGS GYON
- 0F3B ࠉ TIBETAN MARK GUG RTAGS GYAS  
• brackets
- 0F3C ࠊ TIBETAN MARK ANG KHANG GYON
- 0F3D ࠋ TIBETAN MARK ANG KHANG GYAS  
• used for bracketing with a roof over

### Astrological signs

- 0F3E ࠌ TIBETAN SIGN YAR TSHES
- 0F3F ࠍ TIBETAN SIGN MAR TSHES  
• marks which combine with digits

### Consonants

- 0F40 ࠎ TIBETAN LETTER KA
- 0F41 ࠏ TIBETAN LETTER KHA
- 0F42 ࠐ TIBETAN LETTER GA
- 0F43 ࠑ TIBETAN LETTER GHA  
≡ 0F42 ࠐ 0FB7 ࠒ
- 0F44 ࠒ TIBETAN LETTER NGA
- 0F45 ࠓ TIBETAN LETTER CA

## Tibetan

- 0F5C ࠔ TIBETAN LETTER DZHA  
≡ 0F5B ࠓ 0FB7 ࠒ
- 0F5D ࠕ TIBETAN LETTER WA
- 0F5E ࠖ TIBETAN LETTER ZHA
- 0F5F ࠗ TIBETAN LETTER ZA
- 0F60 ࠘ TIBETAN LETTER -A
- 0F61 ࠙ TIBETAN LETTER YA
- 0F62 ࠚ TIBETAN LETTER RA  
• when followed by a subjoined letter = ra mgo
- 0F63 ࠛ TIBETAN LETTER LA
- 0F64 ࠜ TIBETAN LETTER SHA
- 0F65 ࠝ TIBETAN LETTER SSA  
= reversed sha
- 0F66 ࠞ TIBETAN LETTER SA
- 0F67 ࠟ TIBETAN LETTER HA
- 0F68 ࠠ TIBETAN LETTER A  
• base for dependent vowels
- 0F69 ࠡ TIBETAN LETTER KSSA  
≡ 0F40 ࠎ 0FB5 ࠒ
- 0F6A ࠣ TIBETAN LETTER FIXED-FORM RA  
• used only in transliteration and transcription

### Dependent vowel signs

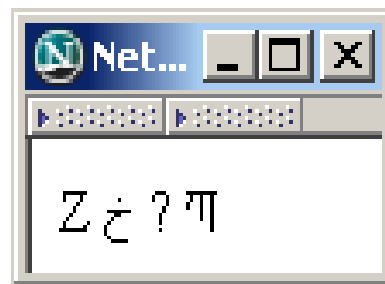
- 0F71 ࠤ TIBETAN VOWEL SIGN AA  
= a-chung  
• common, vowel-lengthening mark
- 0F72 ࠥ TIBETAN VOWEL SIGN I
- 0F73 ࠦ TIBETAN VOWEL SIGN II  
• use of this character is discouraged

## 0F7E

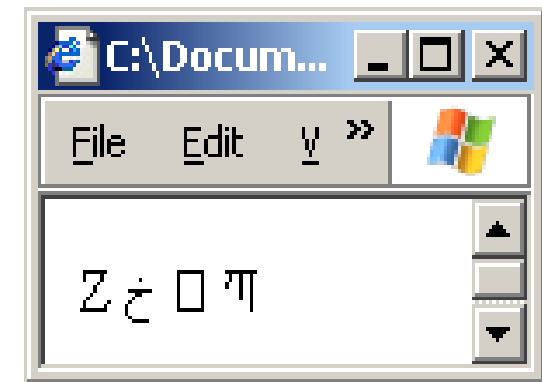
- Das Document Character Set in HTML (ab HTML 4.0)
  - Menge der Zeichen, die in einem HTML Dokument verwendet werden können
  - UNICODE / ISO 10646 („Universal Character Set“, UCS) ist das Document Character Set in HTML, Version 3: >95000 Zeichen
  - Enthält (als Ziel) alle Zeichen der Welt
  - Numerische Zeichenkürzel
    - in HTML `&#90;`;
    - HTML und XML: `&#xFE5;`;
    - CSS: `\fea5`
    - Identifizieren ein Zeichen anhand des CCS, in HTML also immer UNICODE
  - HTML Dokumente sind immer in Unicode, egal wie sie später transferiert werden

- Nicht alle Zeichen müssen darstellbar sein
  - Weil nicht auf lokalen Zeichencode abbildbar
  - Weil keine passende Schrift vorliegt
  - ...
- Vorgehen
  - Nicht vorgeschrieben
  - Teilweise außerhalb der Kontrolle des Darstellers
  - Empfehlung: Visuell klar auf fehlendes Zeichen hinweisen
  - `&#90;`; `&#xFE5;`; `&#10437;`; `&#3904;`;

Netscape 7:



IE 5.5:



- HTML Spezifikation arbeiten immer auf UNICODE Basis
- UNICODE muss nicht benutzt werden, Software muss aber so tun als benutze sie UNICODE
  - Klare Spezifikationen
  - Erlaubt Internationalisierung
  - Unterstützt Lokalisierung
  - Rückwärtskompatibel (ISO 8895-1 gleich unterstem UNICODE Zeichencode)
  - Abstrahiert von Repräsentation der Zeichen in Byteströmen
- UNICODE ist Abstraktionslevel
  - Abstrahiert in der Spezifikation von interner Zeichenkodierung
  - Abstrahiert von Transportrepräsentation

# Zeicheneigenschaften

- Zeichenkodierung (Encoding)
  - Character Encoding Form (CEF)
    - Abbildung einer Zeichenfolge auf Strom gleichgroßer Codes
    - z.B. 

005A	FEA5
------	------
  - Character Encoding Scheme (CES)
    - Abbildung einer Zeichenfolge auf einen Bytestrom
    - z.B. 

5A	00	A5	FE
----	----	----	----
- Zeichensatz
  - Bedeutung unklar, kann Repertoire, Code oder Kodierung meinen
- „charset“
  - meint Encoding!

- Zeichenkodierung
  - Encoding von HTML Dokumenten ist Autoren überlassen
  - Betrifft den Datenstrom zwischen Server und Klienten
  - Klient und Server können die Kodierung aushandeln
  - (Character Set von HTML Dokumenten ist *nicht* verhandelbar)
  - Server und Proxies können Encoding ändern (Transcoding) um Anforderungen des Klienten zu erfüllen
  - Aber: Encoding muss korrekt markiert sein

- Markierung des Encodings
  - HTTP Header Content-Type: `text/html ; charset=EUC-JP`
  - HTML Vorgabe `<meta http-equiv="Content-Type" content="text/html ; charset=EUC-JP" >`
    - `<meta>` so früh wie möglich im Dokument, bis dahin ASCII
    - charset darf bei Transcoding nicht verändert werden
  - charset Attribut bei HTML Elementen
  - XML Vorgabe `<?xml version="1.0" encoding="UTF-8" ?>`
  - CSS2 Vorgabe `@charset "ISO-8859-1";`



# Encoding in HTTP

- Browser kann Präferenzen im HTTP-Request mitteilen:  
GET / HTTP/1.1  
AcceptCharset: i so-8859-1, utf-8; q=0.75, \*; q=0.5
- q gibt Priorität an, \* ist Platzhalter
- Vom Browser abhängig:
  - Microsoft IE: Keine Angabe
  - Netscape 4.72: i so-8859-1, \*, utf-8
  - NS 6.2: I S0-8859-1, utf-8; q=0.66, \*; q=0.66
  - Opera 6.0:  
wi ndows-1252; q=1.0, utf-8; q=1.0, utf-16; q=1.0,  
i so-8859-1; q=0.6, \*; q=0.1

# Encoding in HTTP

---

- Server teilt Encoding in Antwort mit  
200 OK HTTP/1.1  
Content-Type: text/html ; charset=iso-8859-1

<html ...

...

- ISO-8859-1 als Default bei fehlendem charset vorgesehen
- Praktisch nicht haltbar, weil fehlendes charset andere Ursache haben kann
- Browser muss Encoding bei Darstellung auf System abbilden

# Encoding in HTTP

---

- Bei Formulareingaben entstehen Zeichen, die vom Browser an den Server geschickt werden
- Was ist das Encoding der Eingaben auf dem Weg zum Server?
  - Das Encoding der Seite auf der das Formular stand
  - accept-charset Attribut beim Formular:  
`<form accept-charset="ISO-8859-1, utf-8">`

# Transferencoding in HTTP

---

- Zusätzliche Transferencoding verändert den Inhalt einer übermittelten Information
- Beispiel: Kompimierung durch gzip-Verfahren
- In der Anfrage  
GET / HTTP/1.1  
Accept-Encoding: compress; q=0.5, gzip; q=1.0
- In der Antwort  
200 OK HTTP/1.1  
Content-Encoding: gzip
- Kann auf Transportweg (Proxies) geändert werden

# Content Negotiation

- Auswahl passender Information bezüglich der Dimensionen
  - Medienart (Accept: text/html, text/plain)
  - Sprache (AcceptLanguage: en-us; q=0.75, en; q=0.5; \*; q=0.25)
  - Encoding (Accept-Encoding: compress; q=0.5, gzip; q=1.0)
  - Charset (AcceptCharset: iso-8859-1, utf-8; q=0.75, \*; q=0.5)
  - Angegebene Qualitätsmaße
- Server-abhängige Implementierungen
  - z.B. Schema über Dateinamen:
    - foo.en.html
    - foo.html.en
    - foo.en.html.gz

# Zusammenfassung

---

- Internationalisierung und Lokalisierung führen zu lokal anpassbaren und angepassten Diensten und Produkten
- Zunehmend bieten
  - HTML/XML
  - CSS
  - HTTP

## Möglichkeiten zur Internationalisierung und Lokalisierung

- Encoding durch charset Parameter und Header
- Sprache durch lang Attribute und Header
- Transferencoding
- Content-Negotiation

- Tex Texin, Yves Savourel. *Tutorial Standards and Practice Web Internationalization*. 2002.  
<http://www.xencraft.com/resources/webi18ntutorial.pdf>
- The Unicode Consortium. *The Unicode Standard, Version 3.0*, Reading, MA, Addison-Wesley Developers Press, 2000.  
<http://www.unicode.org/standard/standard.html>
- H. Alvestrand. Tags for the Identification of Languages. RFC 3066. 2001.  
<http://www.ietf.org/rfc/rfc3066.txt?number=3066>
- W3C. *Ruby Annotation*. W3C Recommendation 31 May 2001.  
<http://www.w3.org/TR/ruby>
- R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee. RFC 2616 Hypertext Transfer Protocol - HTTP/1.1. June 1999 <ftp://ftp.isi.edu/in-notes/rfc2616.txt>
- Apache HTTP Server Documentation Project. *Apache HTTP Server Version 2.0, Content Negotiation*.  
<http://httpd.apache.org/docs-2.0/content-negotiation.html>

- Internationalization Tag Set (ITS) Version 1.0
  - Zweck: Elemente und Attribute zur Unterstützung von Internationalisierung und Lokalisierung
  - Status: W3C Recommendation 03 April 2007
  - Quelle: <http://www.w3.org/TR/its/>



- Was soll eigentlich lokalisiert werden?

```
<methodResponse>
  <fault>
    <value>
      <struct>
        <member>
          <name>faultCode</name>
          <value><int>4</int></value>
        </member>
        <member>
          <name>faultString</name>
          <value><string>Too many parameters.</string></value>
        </member>
      </struct>
    </value>
  </fault>
</methodResponse>
```

[<http://www.xmlrpc.com/spec>]

# Beispiel: Übersetzung verhindern (Translate category)

- Beim Element vermerkt:

```
<dbk:article xmlns:its="http://www.w3.org/2005/11/its"
  xmlns:dbk="http://docbook.org/ns/docbook"
  its:version="1.0" version="5.0" xml:lang="en">
  <dbk:info>
  <dbk:title>An example article</dbk:title>
  <dbk:author
    its:translate="no">
  <dbk:personname>
  <dbk:firstname>John</dbk:firstname>
  <dbk:surname>Doe</dbk:surname>
  </dbk:personname>
  <dbk:affiliation>
  <dbk:address>
  <dbk:email>foo@example.com</dbk:email>
  </dbk:address>
  </dbk:affiliation>
  </dbk:author>
  </dbk:info>
  <dbk:para>This is a short article.</dbk:para>
</dbk:article>
```

[Beispiele aus Standard]

# Beispiel: Übersetzung verhindern

- Durch globale Regeln bestimmt

```

<myTopic
  xmlns:its="http://www.w3.org/2005/11/its"
  xmlns="myNamespaceURI" id="topic01" xml:lang="en-us">
<prolog>
  <title>Using ITS</title>
  <its:rules version="1.0">
    <its:translateRule selector="//n:term" translate="no"/>
  </its:rules>
</prolog>
<body>
  <p>ITS defines <term>data category</term> as an abstract
  concept for a particular type of information for
  internationalization and localization of XML schemas and
  documents.</p>
</body>
</myTopic>

```

- Hinweise an den Übersetzer, Kontexterläuterung, Auflösung von Mehrdeutigkeiten etc.

```
<myRes
  xmlns:its="http://www.w3.org/2005/11/its" >
  <head>
    <its:rules version="1.0"
      its:translate="no" >
      <its:locNoteRule locNoteType="alert"
        selector="//msg[@id='DisableInfo']">
        <its:locNote>The variable {0} has three possible values: 'printer',
          'stacker' and 'stapler options'.</its:locNote>
      </its:locNoteRule>
    </its:rules>
  </head>
  <body>
    <msg id="DisableInfo">The {0} has been disabled.</msg>
  </body>
</myRes>
```

- Definitionen, Begriffserklärungen
- Beispiel RFC 3530, NFS:

## 1.5. General Definitions

The following definitions are provided for the purpose of providing an appropriate context for the reader.

**Client** The "client" is the entity that accesses the NFS server's resources. The client may be an application which contains the logic to access the NFS server directly. The client may also be the traditional operating system client remote filesystem services for a set of applications.

In the case of file locking the client is the entity that maintains a set of locks on behalf of one or more applications. This client is responsible for crash or failure recovery for those locks it manages.

Note that multiple clients may share the same transport and multiple clients may exist on the same network node.

**Clientid** A 64-bit quantity used as a unique, short-hand reference to a client supplied Verifier and ID. The server is

- Definitionen, Begriffserklärungen

```
<text
```

```
  xmlns:its="http://www.w3.org/2005/11/its" >
```

```
  <its:rules version="1.0" >
```

```
    <its:termRule selector="//term" term="yes"  
      termInfoPointer="id(@def)"/>
```

```
  </its:rules >
```

```
  <p>We may define <term def="TDPV">discoursal point of  
  view</term>
```

```
  as <gloss xml:id="TDPV">the relationship, expressed  
  through discourse
```

```
  structure, between the implied author or some other  
  addresser,
```

```
  and the fiction.</gloss>
```

```
</p>
```

```
</text >
```

- Schriftrichtung
  - Angelehnt an CSS Modell

```
<text
```

```
  xmlns:its="http://www.w3.org/2005/11/its"  
  xml:lang="en" its:version="1.0" >
```

```
<body>
```

```
<par>In Arabic, the title <quote xml:lang="ar"  
  its:dir="rtl">نشاط التدويل، W3C</quote>
```

```
  means <quote>Internationalization Activity,  
  W3C</quote>.</par>
```

```
</body>
```

```
</text>
```

- Zugeordneter Text

```
<text
```

```
  xmlns:its="http://www.w3.org/2005/11/its" >
```

```
<head> ...
```

```
  <its:rules version="1.0" >
```

```
    <its:rubyRule selector="/text/body/img[1]/@alt" >
```

```
      <its:rubyText>World Wide Web Consortium</its:rubyText>
```

```
    </its:rubyRule>
```

```
  </its:rules>
```

```
</head>
```

```
<body>
```

```
   ...
```

```
</body>
```

```
</text>
```



- Festlegung, welches Element die Sprachlichkeit festlegt
- Eigentlich reicht `xml:lang`!

```
<its:rules xmlns:its="http://www.w3.org/2005/11/its"
  version="1.0" >
```

```
  <its:langRule selector="//p"
    langPointer="@mylangattribute"/>
```

```
</its:rules>
```

```
<p mylangattribute="de">Hallo</p>
```

- Ist ein Element Bestandteil des regulären Textes oder ist es ein neuer Textteil
  - Yes:  
<strong>Appaloosa horses</strong> have spotted coats.
  - Nested:  
Palouse horses<fn>A Palouse horse is the same as an Appaloosa.</fn> have spotted coats.
  - No:  
<li>Palouse horses: <p>They have spotted coats.</p> <p>They have been bred by the Nez Perce.</p> </li>

<its:rules

xmlns:its="http://www.w3.org/2005/11/its" version="1.0">

<its:withinTextRule withinText="yes" selector="//b | //em | //i"/>

</its:rules>