Natural Language Processing (NLP) in Real-World Multilingual Production

Stock - Prie

Christian Lieske (Globalization Services, SAP AG)

Seebruck - Chieming

- A Personal View -

Grammatical Framework Summer School (August 2013)

This presentation is purely personal - my employer does not have responsibility for any information contained here.

Overview



NLP in Industry

Part of Solution or Application

(Multilingual) Production



Part of Solution of Application



Multilingual Production – Globalization Tripod



Globalization Size, Impact, and Prospects*

1 g

of online shops only in one language

pages translated

82 % 82 %

1/3 1/3

goes to the translator

2/3

of consumers prefer e-shop in own 1.8 million language

202 million² million

\$ 6.5 billion

revenues for language services market

4500/\$ 450 million \$450 million

words translated

employees/revenue for large Language Service Provider *Numbers not current

Production's Core and Context



Multilingual Production – Challenges (1/4)

Se	Seen from the moon							
-	Internationalize							
-	Localize							
	Translate							





Multilingual Production – Challenges (2/4)



Multilingual Production – Challenges (3/4)



Multilingual Production – Challenges (4/4)

Anyone, anything (proprietary, XML ...), anytime

Scaling, consistency, compliance ...

Coupling

- Object Linking and Embedding, HTTP, Web Services, ...
- Libraries/Application Programming Interfaces/Software Development Kits
- Orchestration (e.g. synchronization of calls, and "bus-like" integration or annotation framework)



http://www.dagstuhl.de/mat/Files/12/12362/12362.LieskeChristian.Slides.pdf

Multilingual Content Processing for Multilingual Production

Ve are presenting on the tcworld conference 2011

Content is more than natural language text

Quality, cost, and delivery count

Often more than just linguistic stuff is in the mix (Natural Language Processing vs. Text Technology)

Sample natural language questions/tasks

Is there existing or new terminology?

Are spelling, grammar, and style alright?

Text technology provides answers not only related to characters, but also to other areas:

Content (e.g. HTML, XML, XML-based vocabularies like DocBook or DITA, ...)Metadata (e.g. RDF)Filters to go from general XML to XLIFF via ITS

Can I recycle an existing translation?

テキストテクノロジーは文字だけではなく、下記の他の分野にも基本になる。

Enablers – Overview (1/2)

You need Natural Language Processing (NLP)/Language Technology (LT) for natural language.

Text Technology is the base for solid, sustainable NLP/LT in real world deployment scenarios.



Enablers – Overview (2/2)

Best Practices and Standardization

Computer-Assisted Linguistic Quality Support

Computer-Assisted Linguistic Assistance

- i. Needs assets
- ii. Creates assets
- iii. Relates to Natural Language Processing

Text Technology Standards for Universal Coverage (1/2)

Think about content with the world in mind

- Can I encode all characters?
- Can my HTML display the content properly?
- Can I translate efficiently?

Only world-ready NLP/LT is solid and sustainable

Unicode standard

- Allows for content creation and processing in a wide range of languages.
- Applied in many contexts (XML, HTML, multilingual Web addresses like <u>http://ja.wikipedia.org/wiki/東京</u>, etc.)

Unicode support should be considered as a key feature of any NLP/LT offering.

Text Technology Standards for Universal Coverage (2/2)

Content formats (e.g. HTML, XML, XML-based vocabularies like DocBook or DITA, ...)

Metadata (e.g. Resource Description Framework)

Filters, e.g. to go from general XML to XLIFF (XML Localization Interchange File Format) based on W3C Internationalization Tag Set (ITS)

. . .

Standards for Universal Efficiency and Effectiveness



Enablers – Canonicalized Content



http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html

Enablers – (Non-intrusive) Universal NLP-related Resource Descriptions

Which parts have to be translated?

Anything I need to know when working on this?



Does the "x" element split a run of text into two linguistic units?



Standards-based Scenario in Main Web Stack



Standards-based Scenario – OKAPI / RAINBOW / CheckMate (1/2)

Okapi Framew	ork			
Welcome				
Okapi	The Okapi Framework is a cross-platform applications that offer extensive support for its goal is to allow tools developers and loca existing ones to best meet their needs, while possible, the project uses and promotes op	and free open-source set of components and localizing and translating documentation and software. alizers to build new localization processes or enhance preserving compatibility and interoperability. Whenever en standards.	Here are some of the tools and a Rainbow — is a GUI appl OmegaT projects, RTF, el verification, translation con mechanism, you can use R CheckMate — is a GUI a Transe Tanger B FE and	pplications based on the framework: lication to launch various utilities related to translation and localization tasks, such as: Text extraction (to XLIFF, tc.) and merging, pre-translation, encoding conversion, terms extraction, file format conversions, quality mparison, search and replace on filtered text, pseudo-translation, and much more. Using the framework's pipeline tainbow to create chains of steps that perform pecific set of tasks specific to your needs. pplication that performs various quality checks on billingual translation files such as XLIFF, TMX, TTX, PO, TS, law dhere filterup domain sunnode Av.
Download the distribution The latest stable r or the latest shape Browse through the dox Read the discussions (Report an issue.	Untitled - Rainbow File View Input Utilities Tools	Help		In to create and maintain segmentation rules. Such rules are used to break down translatable text in more s Okapi's SRX-based segmentation engine. SRX is the Segmentation Rules eXchange format. The application -lows you to see immediately the effects of your segmentation rules on your own sample text, as you edit the rules. boil that offers many functions, including: simple extraction/merging, various file format conversions (TMX, -), access to translation resources, import/export for the Pensieve TM, etc. - Is a plugin to use with OmegaT. It brings transparent support for additional file formats such as TTX, IDML,
 See the list of all the file See the list of the TM an 	Input List 1 Input List 2 Input List 3 Path Relative to the Root	Languages and Encodings Other Settings		fle in OmegaT's pugins directory, restart OmegaT and you are good to go. n server to execute batch processing remotely. Batch configurations which include pre-defined pipelines and executed from Bainbow Longborn provides a BEST interface.
Get the source code. Follow us on Twitter Open source projects can exist Several companies have c End-users (localization ei Brooks Kilne did the Okap		Available steps: Available steps: Raw Document to Filt Filter Events to Raw D Simple Batch Leverag Batch Translation BOM Conversion Used Characters Listin Inline Codes Simplifie Cleanup Segments to Text Uni Create Target Desegmentation Diff Leverage Encoding Conversion	iter Events Document ging Step (Beta) ng il er its Converter	with the project on the Google Code project's People page. Ideas, bug reports, and other contributions to the project.
		Encoder Encoder External Command Extraction Verification Format Conversion Full-Width Conversion Generate SimpleTM GTT Batch Translation	n n n	http://okapi.opentag.com/

Standards-based Scenario – OKAPI / RAINBOW / CheckMate (2/2)



Standards-based Scenario – LanguageTool (1/5)

Homepage News Screenshots Supported Languages Usage Links	LanguageTool is an Open Source proofreading software for English, French, German, Polish, and <u>+more than 20 other</u> languages. It finds many errors that a simple spell checker cannot detect like mixing up <i>there/their</i> and it detects some grammar problems.						
Forum	Try it online						
WikiCheck Development Rule Creator Bug Reports Website Integration	Paste your own text here or check this text too see an few of of the problems that Language	eTool can detecd.					
Java API	Error not found? Improve LanguageTool by writing rules that detect errors.	English 🔽 Check Text					
Javadoc HTTP API HTTP Server Links Wiki	Try LanguageTool without installation, using Java WebStart: + Start LanguageTool (>30 MB, requires Java 6 or later, note: this is not the latest version of L Download	anguageTool)					
Contact	Using LanguageTool locally requires + Java 6 or later. Having problems? Please see the + li Download LanguageTool for LibreOffice/OpenOffice Version 2.2 (34 MB) Download	st of common problems.					
Contact	Using LanguageTool locally requires + Java 6 or later. Having problems? Please see the + li Download LanguageTool for LibreOffice/OpenOffice Version 2.2 (24 MB) Help Help	st of common problems.					
Contact Follow us on twitter Find us on Facebook Get announcements via email	Using LanguageTool locally requires + Java 6 or later. Having problems? Please see the + li Download LanguageTool for LibreOffice/OpenOffice Version 2.2 (34 MB) Help Download LanguageToolFx extension for Mozilla Firefox	st of common problems.					
Contact Follow us on twitter Find us on Facebook Get announcements via email Subscribe	Using LanguageTool locally requires + Java 6 or later. Having problems? Please see the + li Download LanguageTool for LibreOffice/OpenOffice Version 2.2 (24 MB) Help Download LanguageToolFx extension for Mozilla Firefox Check selected text on websites and text in text fields. No Java required!	t of common problems.					

Page last modified: 2013-08-09

http://www.languagetool.org/

Standards-based Scenario – LanguageTool (2/5)

```
<rule id="GENITIV-ARTIKEL">
<pattern>
<token postag_regexp="yes"
postag="SUB:.*"/>
```

```
<token postag_regexp="yes"
postag="ART:(DEF|IND):GEN:.*" skip="-
1"/>
```

```
<token postag_regexp="yes"
postag="SUB:GEN:.*"/>
</pattern>
```

<message>Genitiv gefunden: "<match no="2"/>" Vermeiden Sie den Genitiv.</message>

</rule>

<rule id="GENITIV-POSSESSIVPRONOMEN"> <pattern> <token postag_regexp="yes" postag="SUB:.*"/>

<token postag_regexp="yes" postag="PRO:POS:GEN:.*" skip="-1"/>

<token postag_regexp="yes" postag="SUB:GEN:.*"/>

</pattern>

<message>Genitiv gefunden: "<match no="2"/>" Vermeiden Sie den Genitiv.</message>

</rule>

Courtesy of Annika Nietzio

Standards-based Scenario – LanguageTool (3/5)

1			(mail)		5
🕲 f	orum.openstreetmap.org/viewtopic.php?id=16673	े C	Image: Second Secon	#	E
	Quick post				
	B I U URL IMG CODE	9 9 98	0000000	© 😑 📃	
	QUOIE	Web service	has been used. <u>C</u>	hange settings	
	If you downloads an actual version of the e solved.	d Text languag Mother tong	ge: English (US) Je: German (Germ	nany)	
		1: The prono person form If you downl	un 'you' must be u of a verb: 'downlo oads an actual ver	used with a non-t oad' rsion o	him
		2: Hinweis: " "eigentlich" '(the) latest'	actual/actually" ((Deutsch). Meinte ', 'up-to-date'? oads an actual ver	(Englisch) bedeut en Sie vielleicht 'c rsion of the edit	et u n
	BBCode: on @ [img] tag: on @ Smilies: on	2. Possible s	nolling mistake fe	und	
	Corbert Derview	of the edit	or, the prolbem sh	nould be solved.	

https://addons.mozilla.org/de/firefox/addon/languagetoolfx/

Standards-based Scenario – LanguageTool (4/5)

File Edit View Favorites Tools Help				
LanguageTool Homepage News Screenshots Supported Languages Usage Links	Prüfung auf Leichte Sprache Die Leichte Sprache ist eine besonders leicht verständliche Aus Sprache ausmacht, es gibt zur Orientierung allerdings einige R gegen einige (nicht alle) dieser Regeln zu prüfen. Mehr Inform	drucksweise. Es existiert kein egeln. Mit dieser Seite können ationen zu Leichter Sprache fin	offizieller Standard, was ge Sie LanguageTool benutze Iden Sie beim <mark>Netzwerk Lei</mark>	nau Leichte n, um Texte c <mark>hte Sprach</mark> e
Forum WikiCheck	Fügen Sie hier Ihren Text ein oder benutzen Sie diesen Text a Donaudampfschifffahrt darf da nicht fehlen. Und die Nutzung	als Beispiel, Dieser Text wurde des Genitivs auch nicht. Genitiv gefunden. Vermeiden Sie Hier ignorieren Fehler dieses Typs ignorieren	nur zum Testen geschrieb e den Genitiv.	၅. Die
Forum WikiCheck Development	Fügen Sie hier Ihren Text ein oder benutzen Sie diesen Text a Donaudampfschifffahrt darf da nicht fehlen. Und die Nutzung	als Beispiel, Dieser Text wurde des Genitivs auch nicht. Genitiv gefunden. Vermeiden Sie Hier ignorieren Fehler dieses Typs ignorieren	nur zum Testen geschriebe e den Genitiv.	an. Die

http://www.languagetool.org/de/leichte-sprache/

Standards-based Scenario – LanguageTool (5/5)

ID=39, segment=0:

SAP_TERMINOLOGY_Не траслитерируйте аббревиатуры! Do not transliterate abbreviations! 'IT-оператор'.

T:	'Далее ИТ-о	ператор пл	анирует импорт :	в систему обеспече	ения качес	ва (тестовую сис	тему) всех запрос	ов на перенос,	относящихся	к одному	И ТОМУ Ж	te	
пр	оекту в рамк	ах проекта	технического о	бслуживания SAP Se	olution Mar	ager'							
		A		B		с	D	E F	G	н	1		J
	1 0	Quality Ch	eck Report		*								
ID=	:39, seam 2 1	file:/C:/U	Sort A to Z										fecyc
SA	STYLE 4	ID=1, seg	Sort Z to A										енны
Or a	5 1	ID=1, seg	Sor <u>i</u> by Color										енны
	6	ID=1, seg	Clear Filter Fro	om "(Column B)"									енны
S	The I 7	ID=1, seg	Fitter by Color										енны
SZ	P Solut 8	ID=2, seg	Text Enters										енны
	9	ID=2, segi	Search	• 10									р енны
T: np	Цалее 11 осекту в 12 13	ID=2, segi ID=2, segi ID=3 segi	SAP_GR	аммак_) редложный Аммак_"He" с прича Аммак_"He" с прича	падеж предг стиями пише стиями пише	гся раздельно, при гся раздельно, при гся раздельно, при	едлога (о — 000 — 00; аличии зависимых сл аличии зависимых сл	в – во; при; по; на ов или противопос ов или противопос	: о завершен тавления: Wi тавления: Wi	rite particle - rite particle - rite particle -	ершении, о не-separat не-separat	ely if the	енны
Untitle	d - CheckMat	e								U		~	
e Issu	ues Help												
SAP	Solution M	lanager n	anintonanco n	State									
лее <mark> </mark> дно	ИТ-операто му и тому ж	ор плани ке проек	рует импорт и ту в рамках пр	ојест. в систему обесі роекта техниче	печения ского обо	качества (тесто служивания SA	вую систему) в P Solution Mana	сех запросов ager	на пере	нос, отно	осящихс	+ ^ R +	ки про ки про ки про нетод
лее <mark> </mark> одног AP_TE C:/Use	ИТ-операто му и тому X RMINOLOGY_ rs/D053881/De	ор плани ке проек Не траслит esktop/SM0	рует импорт и гу в рамках пр ерируйте аббрев D1e_RU_Col10/DIT	ојест. в систему обесі роекта техничен иатуры! Do not tran A/mmcpLearningCr	печения ского обо sliterate abl	качества (тесто служивания S/ previations! 'IT-one the_SAP_Solution_	вую систему) в P Solution Mana ратор'. Лапаger_Scenarios_	cex запросов ager for_Application_L	на перен ifecycle_Mar	HOC, OTH	осящихс	+ + + +	си пр си пр си пр цетод цетод тетод тметод
лее <mark> </mark> одног AP_TE C:/Use	ИТ-операто му и тому X RMINOLOGY_ rs/D053881/De Check All	ор плани ке проек Не траслит esktop/SM0	рует импорт и ту в рамках пр ерируйте аббрев D1e_RU_Col10/DIT eck Document	ојест. в систему обесі роекта техниче иатуры! Do not tran A/mmcpLearningCi Configuratior	печения ского обо sliterate abl ontent_Use_	качества (тесто служивания SA previations! 'IT-one the_SAP_Solution_ Session	вую систему) в P Solution Mana ватор'. /anager_Scenarios_ LanguageTool	ccex запросов ager for_Application_L checker warnings	Ha Reper ifecycle_Mar Enabl	HOC, OTH nagement. ed and disa	ОСЯЩИХС dita.ttx abled issue	т Я ^ т s	и пр и пр и пр етод етод метод метод
лее одног AP_TE C:/Use	ИТ-операто му и тому X RMINOLOGY_ rs/D053881/De Check All Text Unit	ор плани ке проек Не траслит esktop/SM0 Сhi Seg	рует импорт и ту в рамках пр ерируйте аббрев D1e_RU_Col10/DIT teck Document Description	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCo Configuratior	печения ского обо sliterate abb ontent_Use_	качества (тесто служивания SA previations! 'IT-one the_SAP_Solution_ Session	вую систему) в P Solution Mana ратор'. Aanager_Scenarios_ (All types of iss Missing target	ccex запросов ager for_Application_L checker warnings sues>	на перен ifecycle_Mar	HOC, OTHO	ОСЯЩИХС dita.ttx abled issue		си при си при си при нетод нетод нетод иметод областа
лее дног AP_TE C:/Use	AT-onepato му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35	р плани ке проек Не траслит esktop/SM0 Сh Seg 0	рует импорт и ту в рамках пр ерируйте аббрев D1e_RU_Col10/DIT tck Document Description Орфографичес	oject. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuratior	печения ского обо sliterate abb ontent_Use_ h	качества (тесто луживания SA previations! 'IT-one the_SAP_Solution_ Session	вую систему) в P Solution Mana batop'. Aanager_Scenarios_ LanguageTool <all iss<br="" of="" types="">Missing target Missing and ext</all>	ccex запросов ager for_Application_L checker warnings sues> tra segments	на перен ifecycle_Mar	HOC, OTHO	ОСЯЩИХС dita.ttx ıbled issue		си при си при си при нетод нетод иметод областа
лее рдног AP_TE C:/Use	AT-onepato му и тому » minology_ rs/D053881/De Check All Text Unit 35 36	р плани ке проек Не траслит esktop/SM0 Сh Seg 0 0	рует импорт и ту в рамках п; ерируйте аббрев D1e_RU_Col10/DIT eck Document Description Орфографичес Орфографичес	oject. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCo Configuratior кая ошибка найден кая ошибка найден	печения ского обо sliterate abb ontent_Use_ h	качества (тесто луживания S/ previations! 'IT-one the_SAP_Solution_ Session	вую систему) в P Solution Mana parop'. Aanager_Scenarios_ (All types of iss Missing target Hissing and ext Empty segment	cex 3anpocoe ager for_Application_L checker warnings sues> tra segments ts	на перен ifecycle_Mar Enabl	HOC, OTHO	ОСЯЩИХС dita.ttx ıbled issue	+ R + * R + *	си пр си пр етод етод етод тетод тметод метод област
лее Dднол AP_TE C:/Use	AT-onepat му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38	нападет п ор плани ке проек esktop/SM0 Сh Seg 0 0 0	рует импорт і гу в рамках пр ерируйте аббрев Die_RU_Coll0/DIT eck Document Description Орфографичес SAP GRAMMAR	oject. в систему обеси роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuration кая ошибка найден 1 Несогласовании	печения ckoro oбo sliterate abl ontent_Use_ h	качества (тест луживания S/ previations! 'IT-one the_SAP_Solution_ Session	Вую систему) в P Solution Mana parop'. Anager_Scenarios_ (All types of iss Missing target Missing and ext Empty segment Target same as ROW White sames di	cex sanpocoe ager for_Application_L checker warnings sues> tra segments ts source fiferences	на перен ifecycle_Mar Enabl	HOC, OTH nagement. ed and disa	ОСЯЩИХС dita.ttx ibled issue referenti		си пр си пр нетод нетод нетод тметод област:
лее AP_TE C:/Use	ИТ-операто му и тому x RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38	нападен т ор плани ке проек не траслип esktop/SM0 Сhи Seg 0 0 0 0	рует импорт и ту в рамках пр ерируйте аббрев DIe_RU_Coll0/DIT eck Document Description Орфографичес SAP_GRAMMAR SAP_GRAMMAR	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCo Configuratior кая ошибка найден кая ошибка найден 1_Несогласованай	печения cкого обо sliterate abb ontent_Use_ ia ia a причастия	качества (тесто луживания S/ previations! 'IT-one the_SAP_Solution_ Session	вую систему) в P Solution Mana parop'. Anager_Scenarios_ All types of iss Missing target Missing and ext Empty segment Target same as son/White spaces di Inline codes dif	cex 3anpocoe ager for_Application_L checker warnings sues> tra segments ts source fferences ferences	на перен ifecycle_Mar Enabl	HOC, OTH nagement. ed and disa	ОСЯЩИХС dita.ttx ibled issue referenti		си пр. си пр. нетод нетод метод област:
лее АР_ТЕ С:/Use	ИТ-операто му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 38 38	нападен т ор плани ке проек He траслип esktop/SM0 Chi Seg 0 0 0 0 1 0	pyet импорт I ту в рамках п ерируйте аббрее DIe_RU_Coll0/DIT eck Document Description Орфографичес SAP_GRAMMAR Орфографичес Орфографичес	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuratior кая ошибка найден 1. Несогласовани кая ошибка найден	печения ckoro oбd sliterate abb ontent_Use_ i ia a a a a a	качества (тесто луживания SA previations! 'IT-one the_SAP_Solution_ Session	Вую систему) в P Solution Mana barop'. Anager_Scenarios_ LanguageTool All types of iss Missing and ext Empty segment Target same as BOM White spaces di Unexpected pat	cex 3anpocoE ager for_Application_L checker warnings sues> tra segments ts source fferences ferences terms	на перен ifecycle_Mai	HOC, OTH nagement. ed and disa ticiple and	ОСЯЩИХС dita.ttx ibled issue referenti		си пр. си пр. нетод нетод етод метод област:
лее днот AP_TE ::/Use 7 7 7 7	AT-oneparc My u TOMy X RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 39 20	нападен т ор плани ке проек He траслит esktop/SM0 Сhi Seg 0 0 0 0 1 0	pyeт импорт I ry в рамках п ерируйте аббрев ole_RU_Coll0/DIT eck Document Description Орфографичес SAP_GRAMMAR Орфографичес сод тЕриниса	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuratior Сопfiguratior кая ошибка найден "1_Несогласованик кая ошибка найден сау ошибка найден	печения ckoro oбd siliterate abl ontent_Use_ i	качества (тесто луживания SA previations! 'IT-one the_SAP_Solution_ Session	BYHO CUCTEMY) B P Solution Mana barop'. Anager_Scenarios_ (All types of iss Missing and ext Empty segment Target same as BON White spaces di Inline codes diff Unexpected pat Suspect pattern	ccex 3anpocoE ager for_Application_L checker warnings sues> tra segments ts source ifferences ferences terms is	на перен ifecycle_Mai	HOC, OTH nagement. ed and disa	осящихс dita.ttx ibled issue referenti		и пр. и пр. и пр. етод етод тметод метод област:
лее Днол AP_TE C:/Use	ИТ-операто му и тому x RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 38 39 39	ор плани ке проек He траслит esktop/SM0 Chr Seg 0 0 0 1 0 0	рует импорт і гу в рамках пр ерируйте аббрев Die_RU_Coll0/DIT teck Document Description Орфографичес SAP_GRAMMAR Орфографичес SAP_TERMINOLI	ојест. в систему обест ооекта техничен иатуры! Do not tran A/mmcpLearningCr (Configuration кая ошибка найден 1_Несогласовании кая ошибка найден ообу_Не траслитери	печения ckoro oбd sliterate abl ontent_Use_ ia ia ia ia ia ia ia ia ia ia	качества (тест луживания S/ previations! 'IT-one the_SAP_Solution_ Session с зависимым сло ревиатуры! Do not	BYIO CUCTEMY) B P Solution Mana parop'. Anager_Scenarios_ (All types of iss Missing target Missing and ext Empty segment Target same as solv White spaces di Inline codes diff Unexpected pattern tran Text length All ruge charged	cex 3anpocoe ager for_Application_L checker warnings sues> tra segments ts source fferences terns s terns terns terns	на перен ifecycle_Maa een par	HOC, OTH nagement. ed and disa	ОСЯЩИХС dita.ttx ibled issue		си пр си пр си пр етод јетод јетод јетод јетод јетод јетод јетод
лее днол AP_TE C:/Use	ИТ-операто му и тому х RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 39 39 39 39	р плани ке проек Не траслит esktop/SM0 Chi Seg 0 0 0 0 1 1 0 0 0	рует импорт і ту в рамках пр ерируйте аббрев Die_RU_Coll0/DIT Description Орфографичес Орфографичес Орфографичес Орфографичес SAP_GRAMMAR SAP_TERMINOL SAP_STYLE_Избо	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuratior кая ошибка найден кая ошибка найден да несогласованик кая ошибка найден кая ошибка найден об9_Не траслитери егайте употреблени	печения cccoro oбc sliterate abb ontent_Use_ на e причастия на на на на на на на на на на на на на	качества (тест луживания S/ previations! 'IT-one the_SAP_Solution_ Session с с зависимым сло ревиатуры! Do not гельных выражен	BY:O CUCTEMY) B P Solution Mana parop'. Anager_Scenarios_ All types of iss Missing target Missing and ext Empty segment Target same as Bow White spaces di Inline codes dif Unexpected pathern Text length tran Text length text length	cex sanpocoe ager for_Application_L checker warnings sues> tra segments ts source fferences ferences terns terns is terns	на перен ifecycle_Mar en par ression:	HOC, OTH nagement. ed and disa ticiple and	осящихс dita.ttx abled issue		си пр си пр си пр нетод нетод тметод метод област
лее днол AP_TE C:/Use 7 7 7 7 7 7 7 7 7 7 7 7 7	ИТ-операто му и тому х :RMINOLOGY_ rs/D053881/De Check All Техt Unit 35 36 38 38 39 39 39 39 39	р плани ке проек Не траслип esktop/SM0 Ch Seg 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	рует импорт и ту в рамках пр ерируйте аббрев DIE_RU_Coll0/DIT eck Document Description Орфографичес Орфографичес SAP_GRAMMAR Орфографичес SAP_TERNINOL SAP_STYLE_Изб Орфографичес	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCo Configuratior кая ошибка найден 1_Несогласовани кая ошибка найден ОGY_Не траслитери егайте употреблени кая ошибка найден	печения cccoro oбo sliterate abl ontent_Use_ ta ta ta ta ta ta ta ta ta ta ta ta ta	качества (тесто луживания SA previations! 'IT-one the_SAP_Solution_ Session I с зависимым сло ревиатуры! Do not	Było CIACTEMY) B P Solution Mana barop'. Anager_Scenarios_ LanguageTool < All types of iss Missing and ett Barby segment Target same as Botw White spaces di Inline codes dif Unexpected pat Suspect patterm tran Text length # Allowed charac Terminology LanguageTool	ccex 3anpocoe ager for_Application_L checker warnings sues> tra segments ts source ffferences ferences terns is ters checker warnings	на перен ifecycle_Mai	HOC, OTH nagement. ed and disa ticiple and	осящихс dita.ttx abled issue		си пр. си пр. си пр. етод јетод јетод јетод јетод област
AP_TE	ИТ-операто му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 39 39 39 39 39 39 39	р плани ке проек He траслит esktop/SM0 Ch Seg 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	рует импорт I ту в рамках п ерируйте аббрее DIe_RU_Coll0/DIT eck Document Description Орфографичес SAP_GRAMMAR Орфографичес SAP_TERMINOLI SAP_STYLE_Изб Орфографичес Орфографичес	ојест. в систему обест роекта техничен иатуры! Do not tran A/mmcpLearningCr Configuratior кая ошибка найден _1_Несогласованик кая ошибка найден оGY_Не траслитери егайте употребленн кая ошибка найден	печения cccoro oбo sliterate abl ontent_Use_ ia ia e причастия ia ируйте аббр е объясни ia ia	качества (тесто луживания SA previations! 'IT-one the_SAP_Solution_ Session с зависимым сло ревиатуры! Do not гельных выражен	Było CIACTEMY) B P Solution Mana barop'. Anager_Scenarios_ LanguageTool. < All types of iss' Missing and ext Empty segment Target same as BON White spaces di Inline codes diff Unexpected pat Suspect pattern an Text length M & Allowed charac Terminology LanguageTool	cex 3anpocoE ager for_Application_L checker warnings uses> tra segments ts source fferences ferences ferences terns terns ters checker warnings	на перен ifecycle_Mai	HOC, OTH nagement. ed and disa ticiple and	dita.ttx bled issue	↓ ↓ ↓ ↑ R	и пр и пр и пр егод егод егод тметод метод
AP_TE	ИТ-операто му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 39 39 39 39 39 39 39 39 39 39 39 39 39	нападен т р плани ке проек He траслит esktop/SM0 Chu Seg 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	рует импорт і гу в рамках п; ерируйте аббрев Die_RU_Coll0/DIT teck Document Description Орфографичес SAP_GRAMMAR Орфографичес SAP_STVLE/Js6 Орфографичес Орфографичес Орфографичес Орфографичес	ојест. в систему обест ооекта техничен иатуры! Do not tran A/mmcpLearningCr (Configuration кая ошибка найден д. Несогласовании кая ошибка найден собУ_Не траслитери егайте употреблени кая ошибка найден кая ошибка найден	печения cxoro oбo sliterate abl ontent_Use_ h на на на на на на на на на на на на на	качества (тест луживания S/ previations! 'IT-one the_SAP_Solution_ Session с зависимым сло ревиатуры! Do not гельных выражен	BYHO CHCTEMY) B P Solution Mana barop'. Anager_Scenarios_ (All types of iss Missing target Missing and ext Empty segment Target same as BON White spaces di Inline codes diff Unexpected pat Suspect pattern tran Text length & Allowe charac Terminology LanguageTool	ccex 3anpocoE ager for_Application_L checker warnings sues> tra segments ts source fferences terms terms ters ters checker warnings	на перен ifecycle_Maa Enabl een par	HOC, OTH nagement. ed and disa ticiple and	dita.ttx ibled issue		си пр си пр си пр тетод тетод тетод тметод област:
AP_TE	ИТ-операто му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 39 39 39 39 39 39 39 39 39 39 39 39 39	р плани ке проек He траслит esktop/SM0 Chr Seg 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	рует импорт і ту в рамках пр ерируйте аббрев Die_RU_Coll0/DIT Description Орфографичес Орфографичес Орфографичес SAP_GRAMMAR Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес	OJECT. в систему обест зоекта техничен иатуры! Do not tran A/mmcpLearningCr Configuration кая ошибка найден кая ошибка найден кая ошибка найден сайте употребленн кая ошибка найден кая ошибка найден кая ошибка найден кая ошибка найден	печения ского обо sliterate abb ontent_Use_ 	качества (тест луживания S/ previations! 'IT-one the_SAP_Solution_ Session с с зависимым сло ревиатуры! Do not гельных выражен	BY:O CUCTEMY) B P Solution Mana Parop'. Manager_Scenarios_ Manager_Scenarios_ (All types of iss Missing target Missing target Missing and ext Empty segment Target same as Bow White spaces di Inline codes dif Unexpected path Suspect pathern Text length Mis Allowed charac Terminology LanguageTool	cex sanpocoe ager for_Application_L checker warnings sues> ra segments ts source fferences ferences terns is terns ters ters checker warnings	Ha neper	HOC, OTH magement. ed and disa ticiple and	осящихс dita.ttx abled issue		си пр. си пр. си пр. етод јетод јетод јетод јетод јетод
AP_TE	ИТ-операто му и тому » RMINOLOGY_ rs/D053881/De Check All Text Unit 35 36 38 38 39 39 39 39 39 39 39 39 39 39 39 40 40	р плани ке проек He траслит esktop/SM0 Chi Seg 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	рует импорт и ту в рамках пр ерируйте аббрев Die_RU_Coll0/DIT eck Document Description Орфографичес SAP_GRAMMAR Орфографичес SAP_TERMINOL SAP_STYLE_Изб Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес Орфографичес	OJECT. в систему обест соекта техничен иатуры! Do not tran A/mmcpLearningCo Configuration кая ошибка найден кая ошибка найден	печения ckoro oбd sliterate abl ontent_Use_ ta ta ta ta ta ta ta ta ta ta ta ta ta	качества (тесто луживания S/ previations! 'IT-one the_SAP_Solution_ Session I с зависимым сло ревиатуры! Do not гельных выражен	Było CIACTEMY) B P Solution Mana harop'. Anager_Scenarios_ Anager_Scenarios_ All types of iss Missing and ext Missing and ext	cex 3anpocoe ager for_Application_L checker warnings sues> tra segments source fferences ferences terns s terns ters checker warnings	на перен ifecycle_Mai	HOC, OTH nagement. ed and disa ticiple and	осящихс dita.ttx bbled issue	▼ ↓ ↑ R	и при и при и при и тор нетод нетод нетод метод

Multilingual content processing needs help

"Which data elements need to be processed by NLP?"

<rsrc id="123"> ...

- <data type="text">images/cancel.gif</data>
- <data type="position">12,20</data>
- <data type="text">Cancel</data>
- <data type="position">60,40</data>
- <data type="text">Number of files: </data>

</rsrc>

ITS 2.0 – The help



ITS 2.0 Basic principles

Say important things

About specific content

In a standard way

• With agreed upon syntax and values

1. Say important things: ITS 2.0 "data categories"

Translate, Localization Note, Terminology, Directionality, Language Information, Elements Within Text, Domain, Text Analysis, Locale Filter, Provenance, External Resource, Target Pointer, Id Value, Preserve Space, Localization Quality Issue, Localization Quality Rating, MT Confidence, Allowed Characters, Storage Size

Definition in prose

Selection of content via two approaches

2. About specific content: Content selection approaches

ITS selection can be compared to

- global = "style" element
- local = "style" attribute

Selection global	 <u>XPath</u> to select markup nodes
Selection local	ITS local attributes
<rsrc> <its:rules its<br="" text"="" xmlns:its="https://www.sits:rules.com/sits:translateRu
</its:rules>
•<data type=">•<data <="" th="" type="position"><th>ttp://www.w3.org/2005/11/its" le <u>selector="//data"</u> translate="no"/> s:translate="yes">Cancel</th></data> n">60,40 </its:rules></rsrc>	ttp://www.w3.org/2005/11/its" le <u>selector="//data"</u> translate="no"/> s:translate="yes">Cancel

3. In a standard way (1/2)



3. In a standard way (2/2)



Why ITS <u>2.0</u>? (1/2)

ITS 1.0 = simplified view of multilingual content production

Too limited for comprehensive automated content processing/usage scenarios (see <u>http://www.w3.org/TR/mlw-</u> <u>metadata-us-impl/</u> for various ITS 2.0 usage scenario descriptions)

Example gap: too few data categories

Why ITS 2.0? (2/2)

Coverage for additional types of content: HTML5

- Bridge to Web & app content
- Accommodate relevant HTML5 markup (e.g. HTML5 "translate" attribute behaviour)

Easy mapping/conversion to other formats

- XML Localization Information Markup (XLIFF; status: <u>informal</u> mapping, under discussion) = workflows
- Natural Language Processing Interchange Format (NIF) = bridge to the Semantic Web and Natural Language Processing

Example: MT Confidence

Score from machine translation engine

Example for new ITS capability: Tool traceability

```
<!DOCTYPE html> ...
<body its-annotators-ref="mt-
confidence|file:///tools.xml#T1">
<span its-mt-confidence="0.8982">Dublin is the
capital of Ireland.</span>
</body></html>
```

Example: Locale Filter

Content relevant only for a specific locale

```
<!DOCTYPE html> ...
<div its-locale-filter-list="*-ca">
Text for Canadian locales.
</div>
<div its-locale-filter-list="*-ca" its-locale-filter-
type="exclude">
Text for non-Canadian locales.
</div> ...
```

Example: Localization Quality Issue

For quality assessment

<!DOCTYPE html> ... < its-loc-quality-issue-comment="should be 'quality'" its-loc-quality-issue-profileref=http://example.org/qaMovel/v1 its-loc-quality-issue-severity=50 its-loc-quality-issue-type=spelling>qulaity

° 4



- 1. "Filtering" with Okapi Rainbow (built-in filter)
- 2. "Filtering" with Okapi Rainbow (custom filter configuration based on W3C Internationalization Tag Set)
- 3. Browser-based demo of LanguageTool

Discussion/Ideas/Suggestions

- **1. GF** for checking term entry conventions
- 2. GF for term lookup
- 3. GF in automated pre-editing for MT
- 4. GF in automated post-editing for MT
- 5. GF and markup
- 6. GF from Okapi (e.g. one or more steps)
- 7. Okapi from GF (e.g. as pre- and post-processor)
- 8. GF and LanguageTool
- 9. GF and TermBase eXchange (TBX)
- **10.** GF and Translation Memory eXchange (TMX)
- **11. GF and Pseudo-translation**



More information on W3C ITS:

http://www.w3.org/TR/its/ http://www.w3.org/TR/its20/

<u>http://www.w3.org/International/its/ig/</u> <u>http://lists.w3.org/Archives/Public/public-i18n-its-ig</u> (public list, free to subscribe)

Contact:

Christian Lieske christian.lieske@sap.com www.sap.com



Disclaimer

All product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

This document may contain only intended strategies, developments, and is not intended to be binding upon the authors or their employers to any particular course of business, product strategy, and/or development. The authors or their employers assume no responsibility for errors or omissions in this document. The authors or their employers do not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

The authors or their employers shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The authors have no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.