The World Speaks One Language - Yours ™

Company & Technology Overview

Product and Technology Platform Overview

Kirti Vashee – kirti.vashee@asiaonline.net

Copyright © 2009, Asia Online Pte Ltd

Asia Online Company Objectives

Asia Online™
The World Speaks One Language - Yours™

Revolutionize the Internet experience for non-English speakers in Asia

Revolutionize the enterprise translation process with a comprehensive, continuous learning SMT platform

Provide 1 billion+ local-language pages online using mostly translated open license content, combined with compelling portal and social networking style services in Thailand, Indonesia, India, Malaysia, Philippines, Vietnam and China, Japan & Korea

SaaS environment that allows data cleaning and preparation, develop SMT engines on demand and enable ongoing comprehensive post editing and correction to continuously improve engines

The Consumer Market

The Enterprise Market

Large Buyer & Publisher Perspective

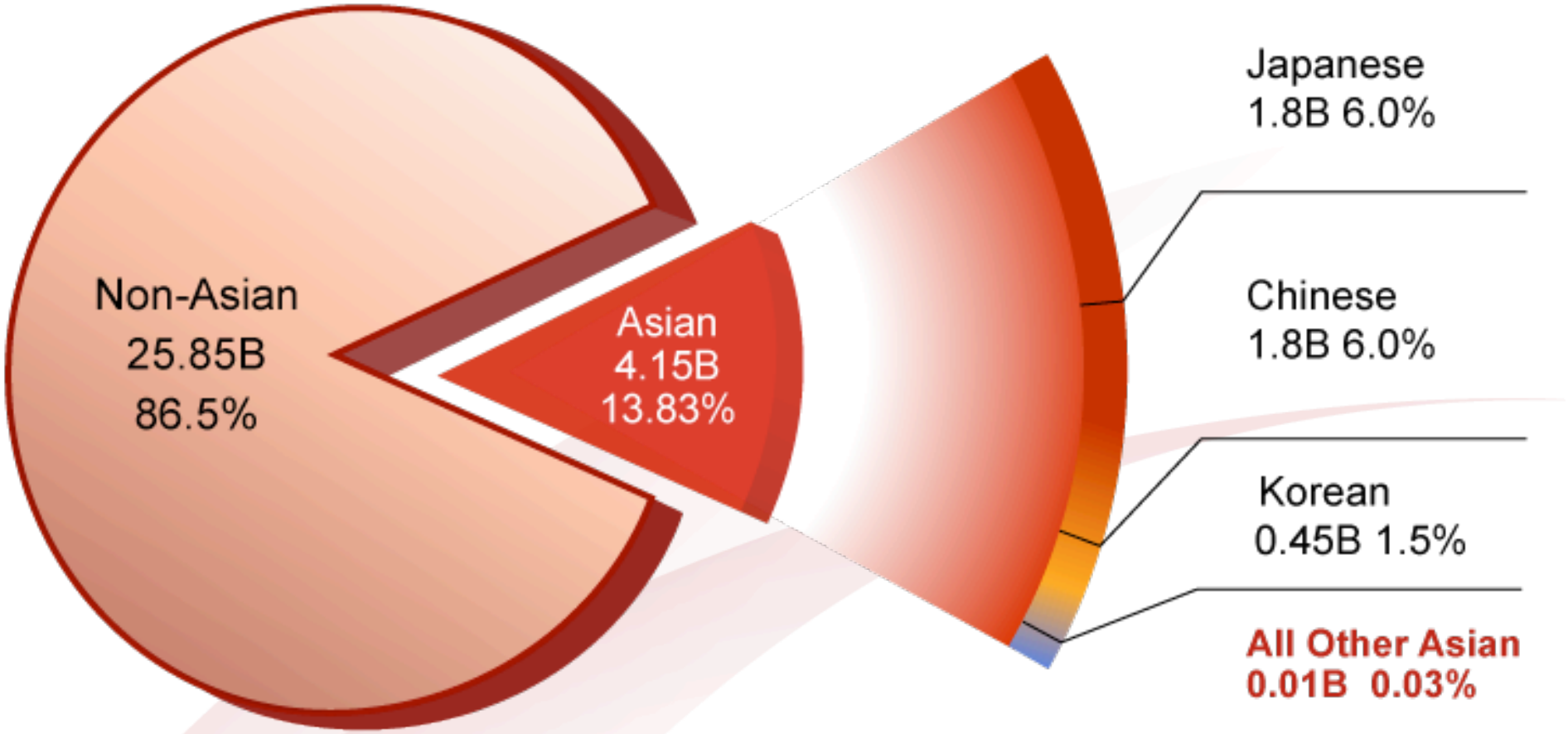
Translation Tools Vendor Perspective

Copyright © 2009, Asia Online Pte Ltd

Asia Online™
The World Speaks One Language - Yours™

Asian Language Content Gap

Information Poverty



Copyright © 2009, Asia Online Pte Ltd

What makes Asia Online unique?

- The only SMT technology provider that is **also a major user** of ALT technology on one of the largest translation projects in the world - English Wikipedia (1B Words+) into 11 Asian languages using MT and crowdsourcing.
- The translation tools and technology platform used to accomplish this, is also made available as a SaaS product and onsite servers for the enterprise translation market.

Constant dialogue between the Users and Developers to ensure that rapid translation quality evolution is a reality

Copyright © 2009, Asia Online Pte Ltd

Asia Online™

The World Speaks One Language - Yours™

Technology Overview

Copyright © 2009, Asia Online Pte Ltd

Asia Online™
The World Speaks One Language - Yours™

The Translation Platform Overview

Data Management

- Bilingual Data Preparation & Cleaning
- Bilingual Data Normalization & Optimization
- Source Cleanup and Preparation
- Grammar and Spelling validation
- Monolingual Data Extraction & Analysis

Hybrid SMT Engine

- SMT System Training & Development
- Monolingual Data Training
- Ongoing Corpus Refinement and Tuning
- Analysis and Evaluation of N-grams

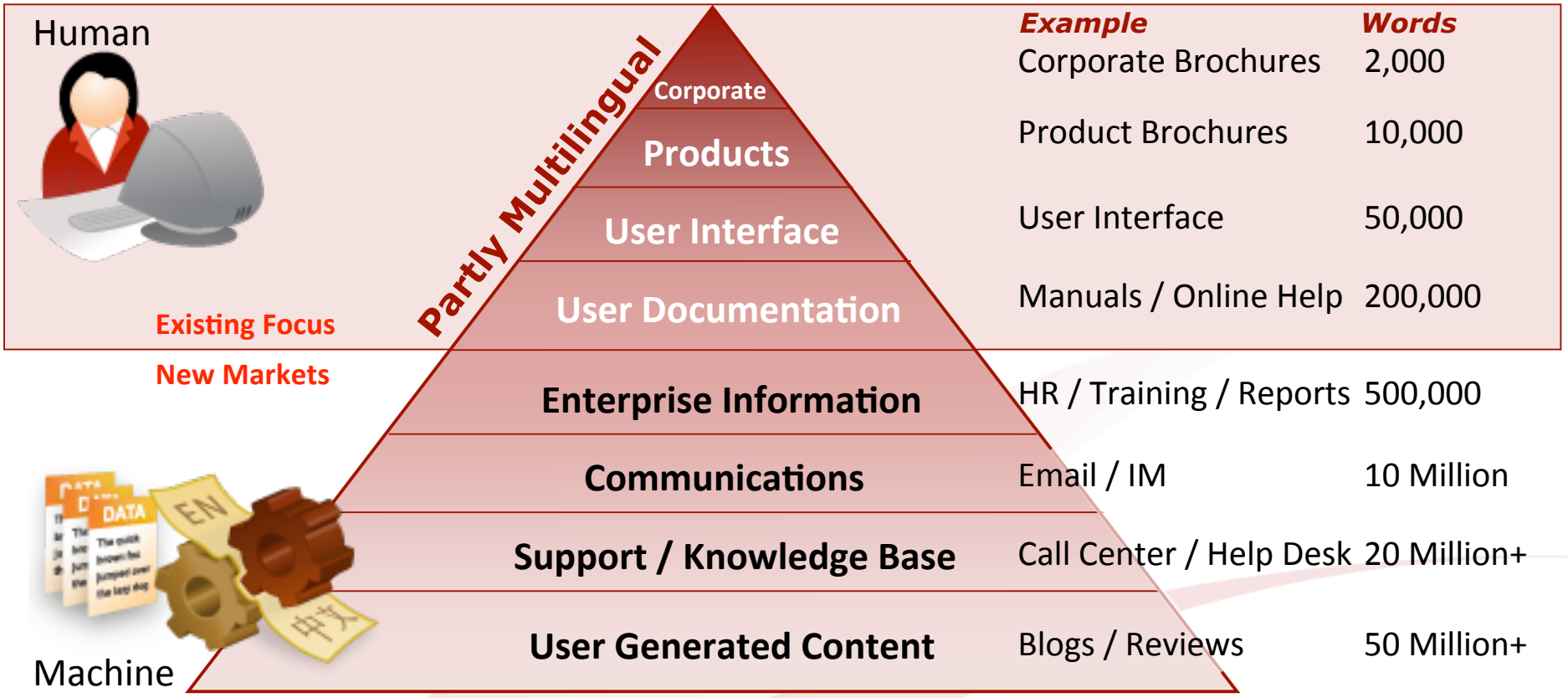
Output Proofing and Editing

- Error Pattern Identification & Correction
- Automated error correction tools
- Continuing Cycle of Exception Identification and Correction
- Development of small sets of new data to correct errors

Copyright © 2009, Asia Online Pte Ltd

Expanding the Reach of Translation

Asia Online™
The World Speaks One Language - Yours™



	Example	*Words*
Corporate	Corporate Brochures	2,000
Products	Product Brochures	10,000
User Interface	User Interface	50,000
User Documentation	Manuals / Online Help	200,000
Enterprise Information	HR / Training / Reports	500,000
Communications	Email / IM	10 Million
Support / Knowledge Base	Call Center / Help Desk	20 Million+
User Generated Content	Blogs / Reviews	50 Million+

Problem: • **Only 0.5% of what needs to be translated today is being translated due to cost and time constraints.**

Solution: • **Machine translation offers , "good enough" quality alternative to human that is faster and lower cost and can deliver huge volumes of content quickly – millions of words per day.**
•Where "gold-standard" translations are required, MT + human post editing can deliver the same content at high quality, significantly faster.

Copyright © 2009, Asia Online Pte Ltd

Connecting the Language Studio™ Suite

Asia Online™
The World Speaks One Language - Yours™

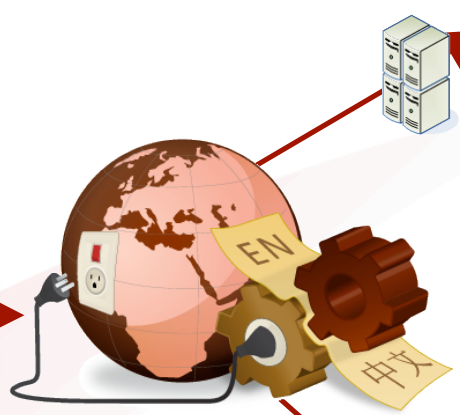
Customer Site

Private Data Storage

Project and Resource Management

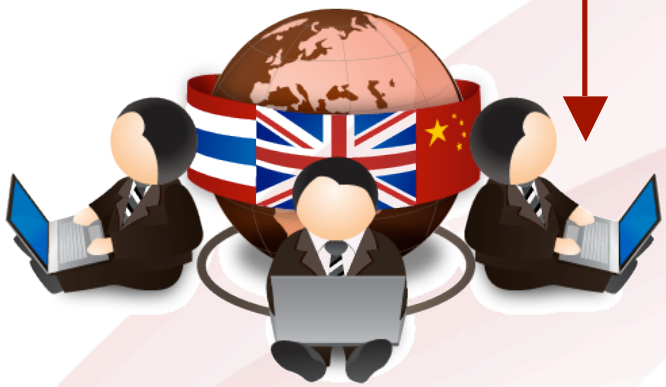
Asia Online Software as a Service

Translation

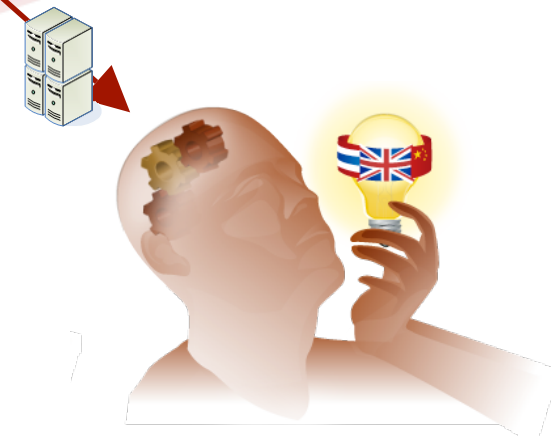


Internet

External Editors and Translators

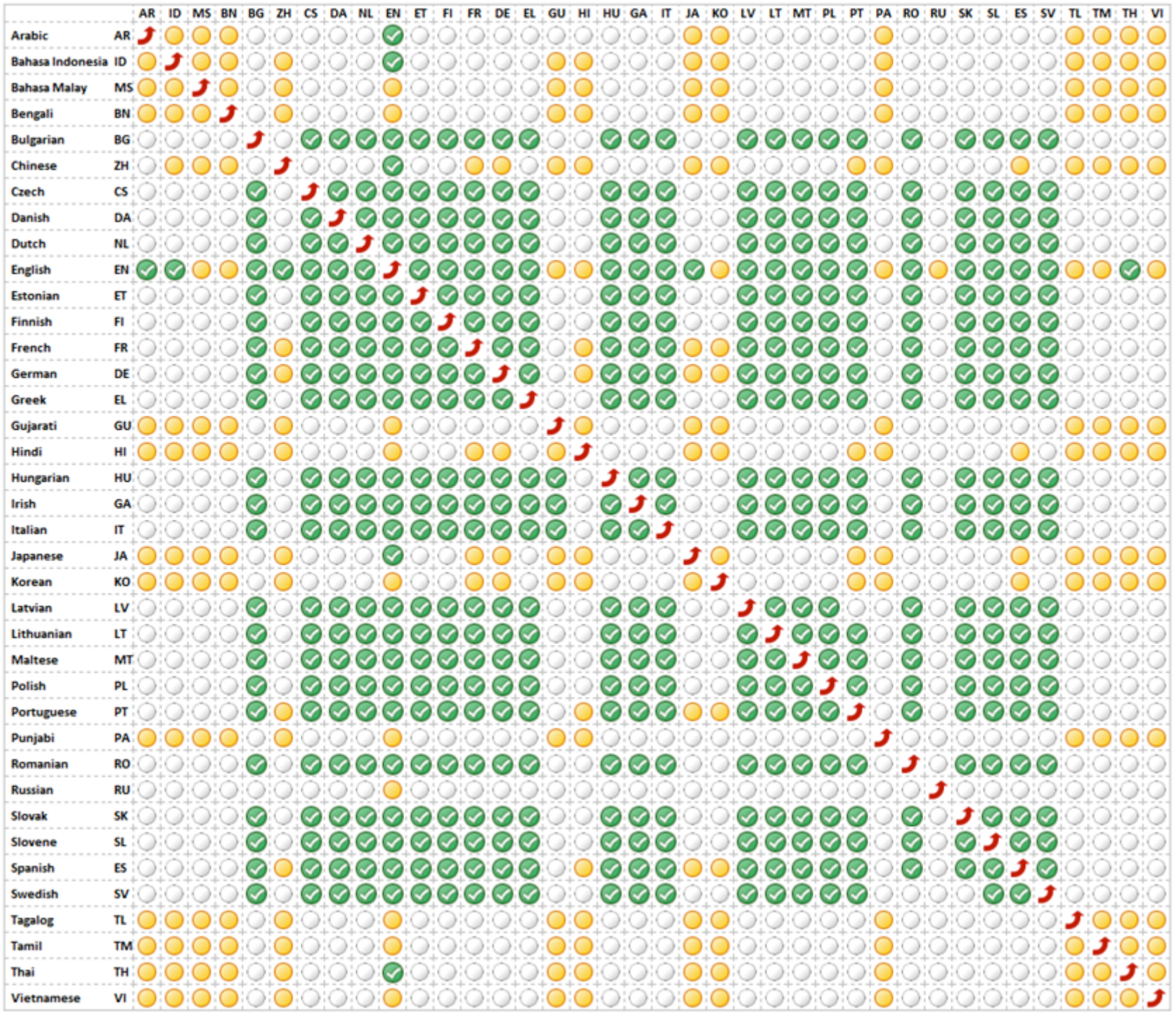


Distributed Post Editing



Training

Copyright © 2009, Asia Online Pte Ltd



Choose Language Pair

Key

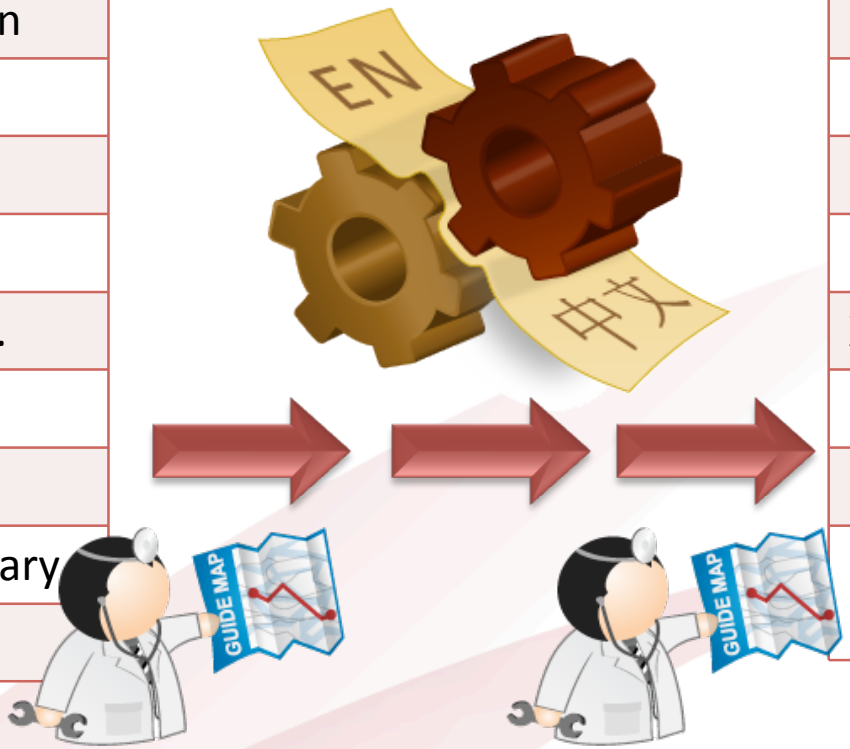
- Available Now (516)
- Planned / Under Development (228)
- Not Planned in near future, but possible based on customer request.

Other language pairs will be created as demand requires or based on customer request.

Understanding the Language Studio™ Hybrid Rules and SMT Engine Model

Pre-Processing Rules
Sentence Segmentation
Word Segmentation
Phrase Reordering
Dates and Numbers
Patterns, Formulas etc.
Pre-Normalization
Spell Checking
Custom Runtime Glossary
Pre-Formatting

Statistical Machine Translation



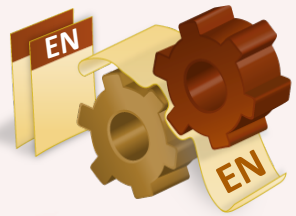
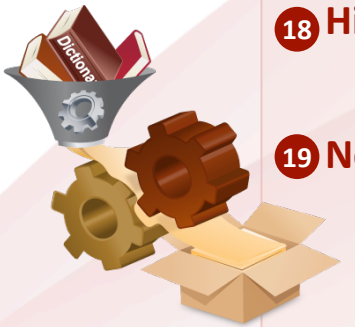
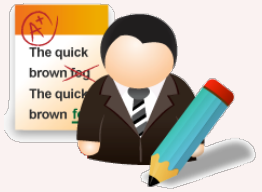
Post-Processing Rules
Capitalization
Post-Formatting
Grammar Checking
Post-Normalization
XML Tag Reinsertion
Currency Conversion
Cross Referencing
Other custom post processing

Pre-Processing and Post-Processing Rules are implemented in a C# script, which can be customized to fit almost any requirements. This is a true hybrid approach, where rules are used to guide the statistical engine to greater accuracy.

Copyright © 2009, Asia Online Pte Ltd

Key Stages in Engine Customization

Asia Online™ The World Speaks One Language - Yours™

Training Data	Preparation	Train	Diagnostic Engine	Train	Version Release →
1 Foundation Language Pair 2 Foundation Domain					
3 Custom Data →	13 Cleaning				
4 Bilingual Data					
5 Glossaries →	14 Data Generation				
6 Phrase Pairs → (TMX, XLIFF, CSV)	15 Validate Alignment and Quality				
7 Bilingual Documents → (MS Word, Plain Text, etc.)	16 Alignment				
8 Monolingual Data			20 Error Pattern Analysis & Correction		
9 URLs →	17 Data mining for grammatical patterns				
10 Plain Text					
11 Tuning Set	Optional (Recommended)		Optional (Recommended)		Optional (Recommended)
12 Test Set	18 High Frequency Term & Phrase Analysis		21 Unknown Word Analysis		24 Post Edit and Train as required
	19 Normalization		22 Normalization		
			23 Error Analysis		

Copyright © 2009, Asia Online Pte Ltd

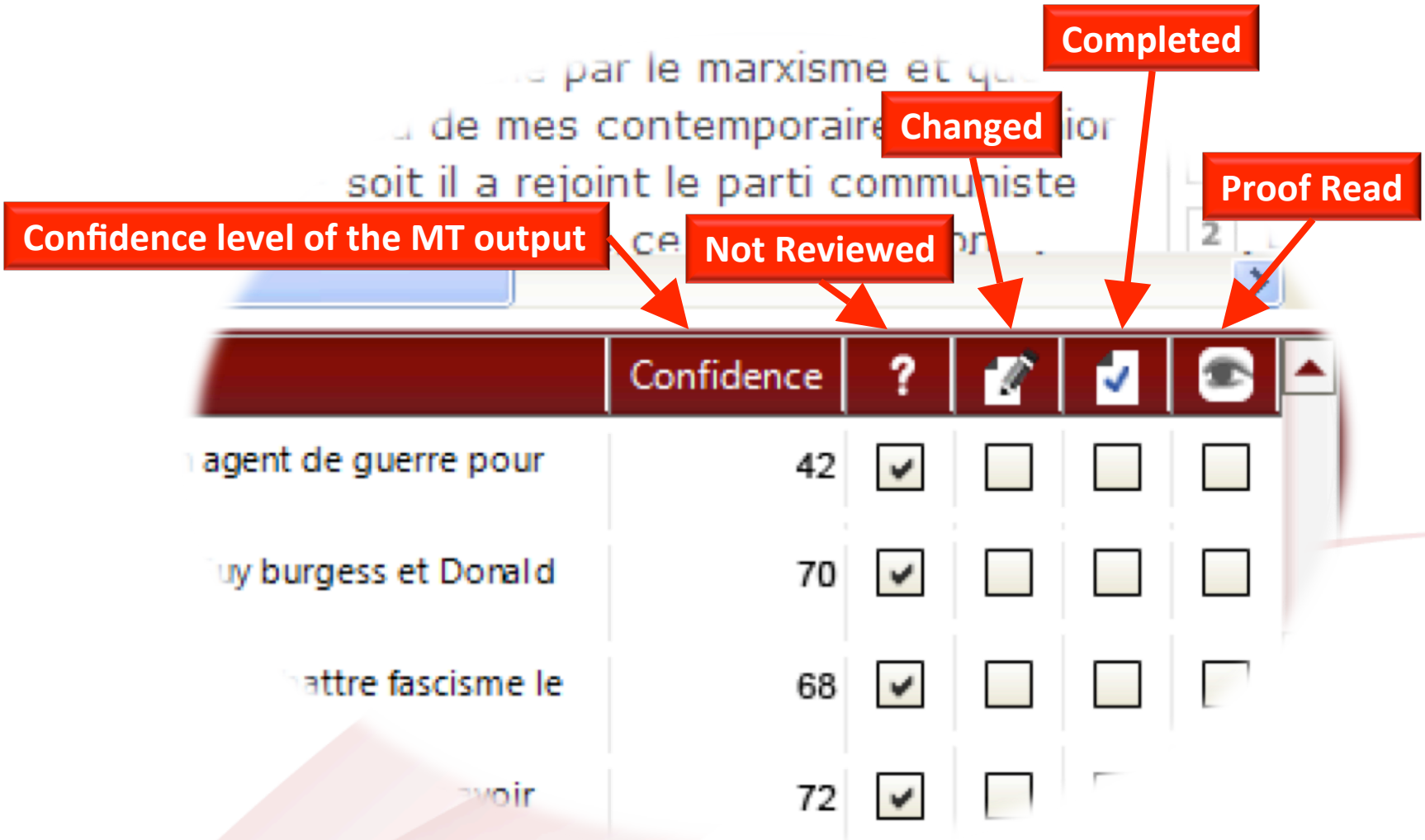
Asia Online™

The World Speaks One Language - Yours™

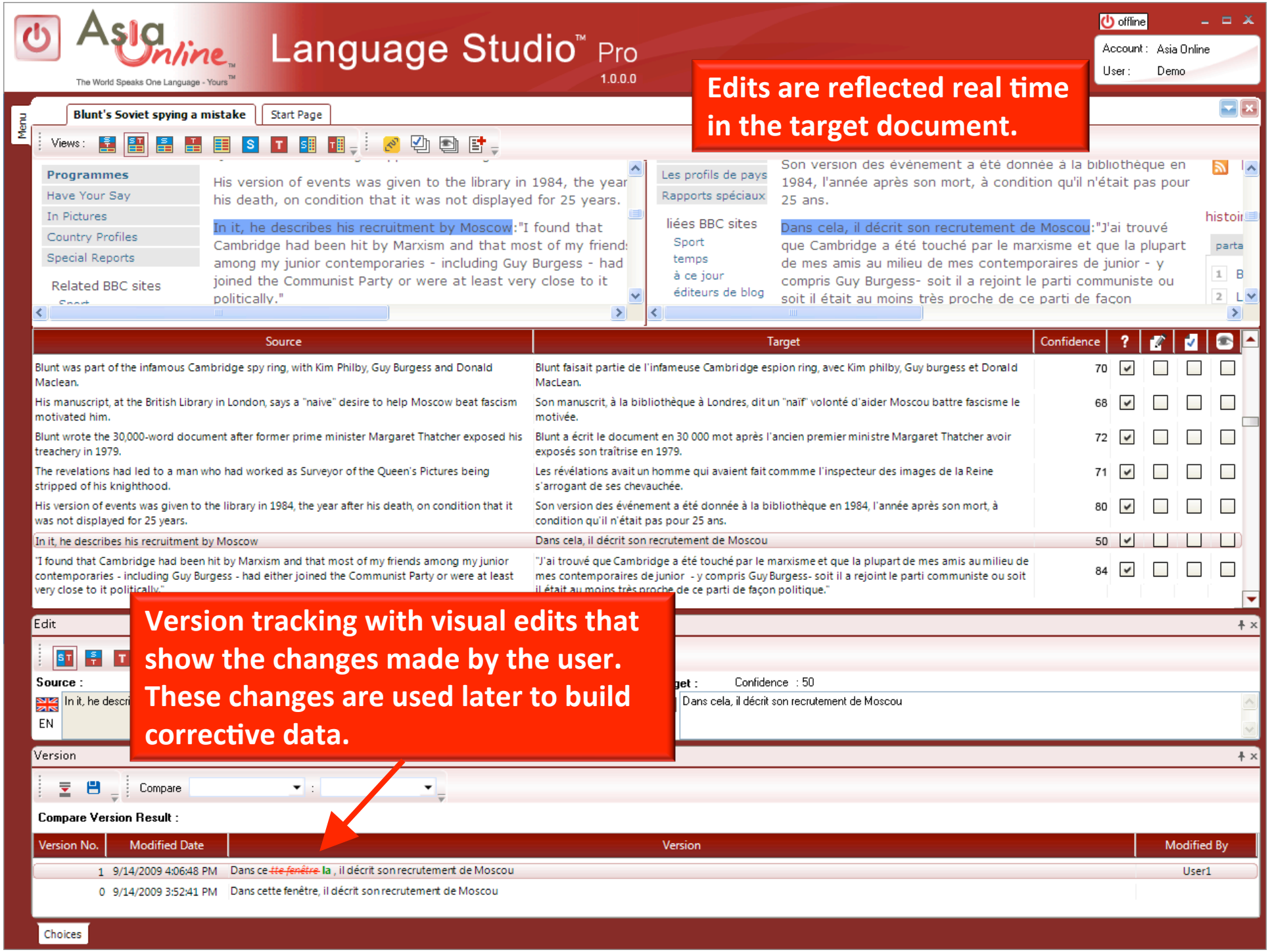
The System Interface & Screenshots

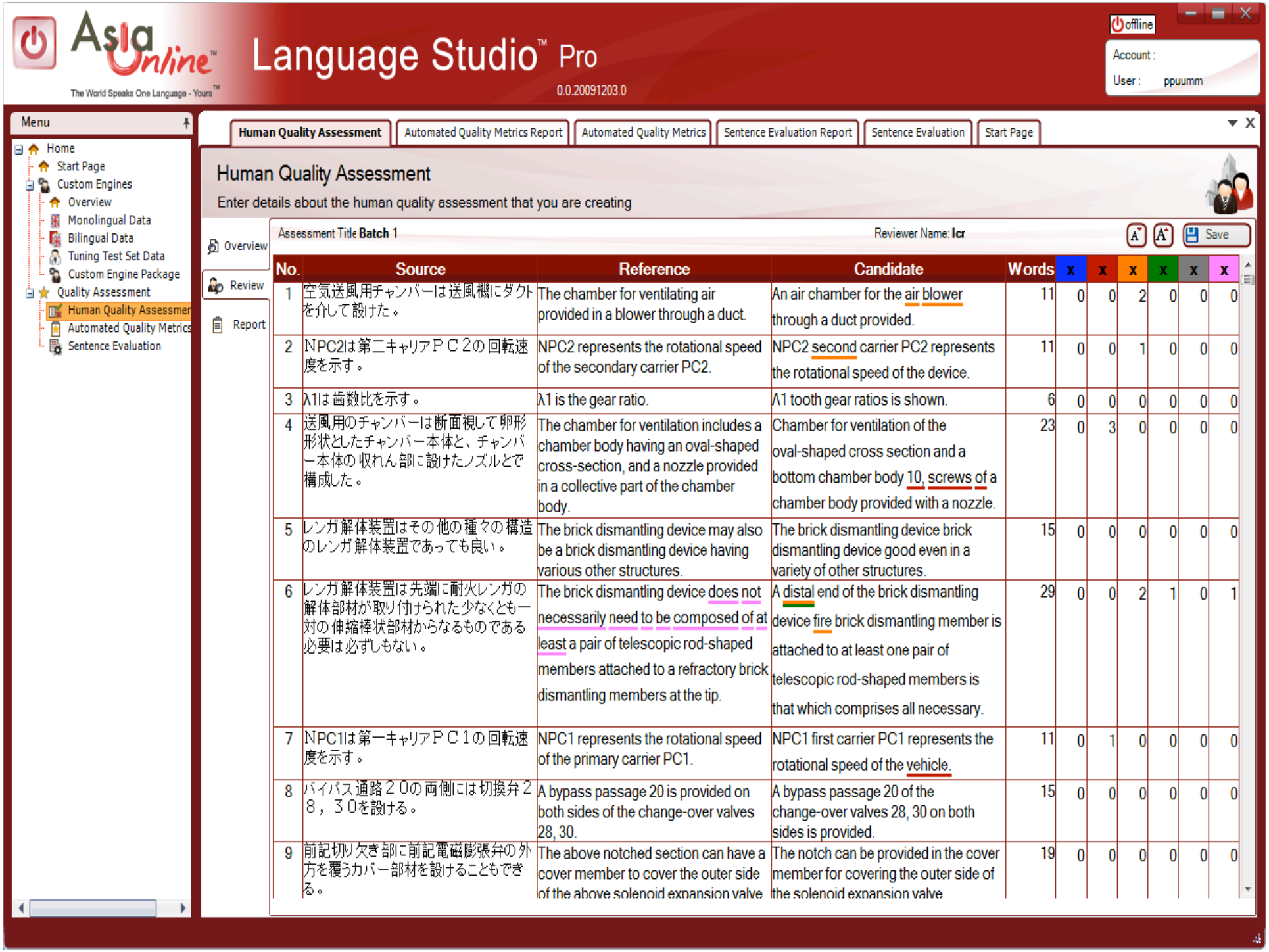
Copyright © 2009, Asia Online Pte Ltd

Asia Online
The World Speaks One Language - Yours
Language Studio™ Pro
1.0.0.0
offline
Account : Asia Online
User : Demo
Menu
Blunt's Soviet spying a mistake
Start Page
Views :
Source
Target
Health
Science & Environment
Technology
Video and Audio
Programmes
Have Your Say
In Pictures
Country Profiles
Special Reports
Related BBC sites
Library in London, says a "naive" desire to help Moscow beat fascism motivated him.
Blunt's memoirs revealed little about his spying activities
Blunt wrote the 30,000-word document after former prime minis Margaret Thatcher exposed his treachery in 1979.
The revelations had led to a man who had worked as Surveyor Queen's Pictures being stripped of his knighthood.
His version of events was given to the library in 1984, the year his death, on condition that it was not displayed for 25 years.
In it, he describes his recruitment by Moscow:"I found that Cambridge had been hit by Marxism and that most of my friend among my junior contemporaries - including Guy Burgess - had joined the Communist Party or were at least very close to it politically."
Aussi la nouvelle
Vous dire
images dans
Les profils de pays
Rapports spéciaux
liées BBC sites
Sport
temps
à ce jour
éditeurs de blog
Blunt a écrit le document en 30 000 mot après l'ancien premier ministre Margaret Thatcher avoir exposés son traîtrise en 1979.
Les révélations avait un homme qui avaient fait commme l'inspecteur des images de la Reine s'arrogant de ses chevauchée.
Son version des événement a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.
Dans cette fenêtre, il décrit son recrutement de Moscou:"J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess- soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de facon
Source
Target
Confidence
He supplied hundreds of secret documents to the Soviets while a wartime agent for MI5.
Il a fourni des centaines de des documents secrets à les Soviets alors qu'un agent de guerre pour MI5 .
42
Blunt was part of the infamous Cambridge spy ring, with Kim Philby, Guy Burgess and Donald Maclean.
Blunt faisait partie de l'infameuse Cambridge espion ring, avec Kim philby, Guy burgess et Donald MacLean.
70
His manuscript, at the British Library in London, says a "naive" desire to help Moscow beat fascism motivated him.
Son manuscrit, à la bibliothèque à Londres, dit un "naïf" volonté d'aider Moscou battre fascisme le motivée.
68
Blunt wrote the 30,000-word document after former prime minister Margaret Thatcher exposed his treachery in 1979.
Blunt a écrit le document en 30 000 mot après l'ancien premier ministre Margaret Thatcher avoir exposés son traîtrise en 1979.
72
The revelations had led to a man who had worked as Surveyor of the Queen's Pictures being stripped of his knighthood.
Les révélations avait un homme qui avaient fait commme l'inspecteur des images de la Reine s'arrogant de ses chevauchée.
71
His version of events was given to the library in 1984, the year after his death, on condition that it was not displayed for 25 years.
Son version des événement a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.
80
In it, he describes his recruitment by Moscow
Dans cette fenêtre, il décrit son recrutement de Moscou
50
"I found that Cambridge had been hit by Marxism and that most of my friends among my junior contemporaries - including Guy Burgess - had either joined the Communist Party or were at least very close to it politically."
"J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess- soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de facon politique."
84
Edit
Source :
EN
In it, he describes his recruitment by Moscow
Target :
Confidence : 50
FR
Dans cette fenêtre, il décrit son recrutement de Moscou
Version
Choices



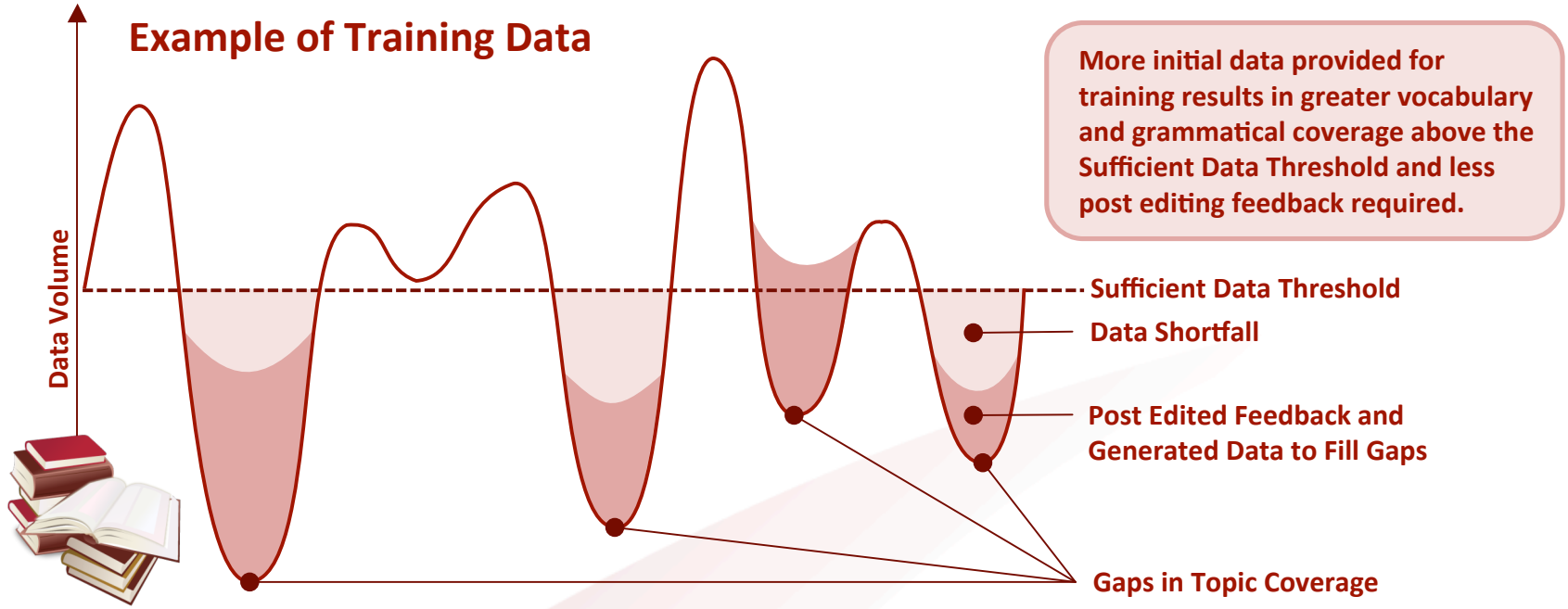
Copyright © 2009, Asia Online Pte Ltd


Asia Online
The World Speaks One Language - Yours
Language Studio™ Pro
1.0.0.0
offline
Account : Asia Online
User : Demo
Edits are reflected real time in the target document.
Blunt's Soviet spying a mistake
Start Page
Menu
Views :
Programmes
Have Your Say
In Pictures
Country Profiles
Special Reports
Related BBC sites
His version of events was given to the library in 1984, the year his death, on condition that it was not displayed for 25 years.
In it, he describes his recruitment by Moscow:"I found that Cambridge had been hit by Marxism and that most of my friends among my junior contemporaries - including Guy Burgess - had joined the Communist Party or were at least very close to it politically."
Les profils de pays
Rapports spéciaux
liées BBC sites
Sport
temps
à ce jour
éditeurs de blog
Son version des événement a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.
Dans cela, il décrit son recrutement de Moscou:"J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess- soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de façon
Source
Target
Confidence
Blunt was part of the infamous Cambridge spy ring, with Kim Philby, Guy Burgess and Donald Maclean.
Blunt faisait partie de l'infameuse Cambridge espion ring, avec Kim philby, Guy burgess et Donald MacLean.
70
His manuscript, at the British Library in London, says a "naive" desire to help Moscow beat fascism motivated him.
Son manuscrit, à la bibliothèque à Londres, dit un "naïf" volonté d'aider Moscou battre fascisme le motivée.
68
Blunt wrote the 30,000-word document after former prime minister Margaret Thatcher exposed his treachery in 1979.
Blunt a écrit le document en 30 000 mot après l'ancien premier ministre Margaret Thatcher avoir exposés son traîtrise en 1979.
72
The revelations had led to a man who had worked as Surveyor of the Queen's Pictures being stripped of his knighthood.
Les révélations avait un homme qui avaient fait commme l'inspecteur des images de la Reine s'arrogant de ses chevauchée.
71
His version of events was given to the library in 1984, the year after his death, on condition that it was not displayed for 25 years.
Son version des événement a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.
80
In it, he describes his recruitment by Moscow
Dans cela, il décrit son recrutement de Moscou
50
"I found that Cambridge had been hit by Marxism and that most of my friends among my junior contemporaries - including Guy Burgess - had either joined the Communist Party or were at least very close to it politically."
"J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess- soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de façon politique."
84
Version tracking with visual edits that show the changes made by the user. These changes are used later to build corrective data.
Edit
Source :
EN
Confidence : 50
Dans cela, il décrit son recrutement de Moscou
Version
Compare
Compare Version Result :
Version No.
Modified Date
Version
Modified By
1 9/14/2009 4:06:48 PM
Dans ce ~~tte fenêtre~~ la , il décrit son recrutement de Moscou
User1
0 9/14/2009 3:52:41 PM
Dans cette fenêtre, il décrit son recrutement de Moscou
Choices



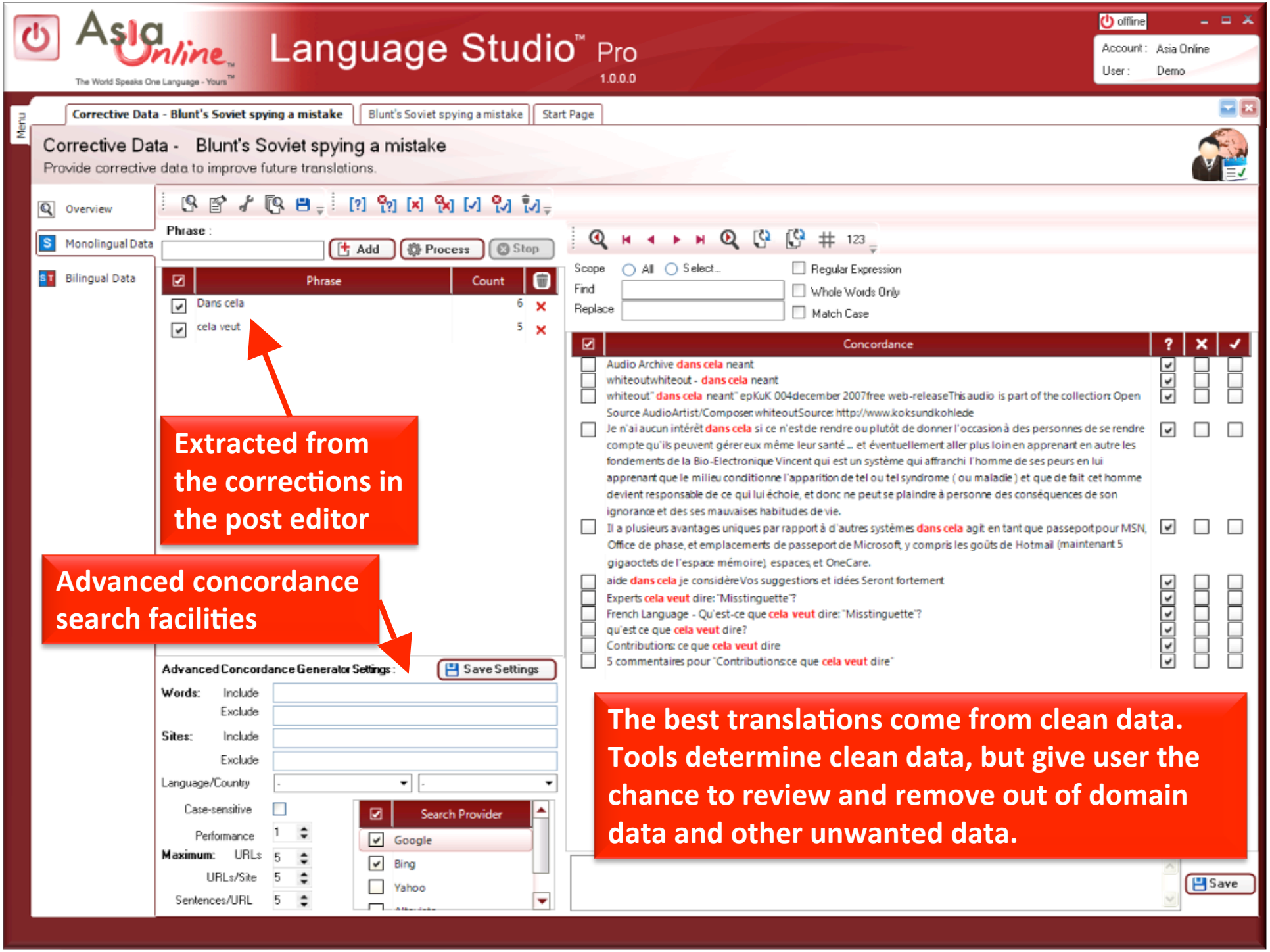
No.	Source	Reference	Candidate	Words	x	x	x	x	x	x
1	空気送風用チャンバーは送風機にダクトを介して設けた。	The chamber for ventilating air provided in a blower through a duct.	An air chamber for the air blower through a duct provided.	11	0	0	2	0	0	0
2	NPC2は第二キャリアＰＣ２の回転速度を示す。	NPC2 represents the rotational speed of the secondary carrier PC2.	NPC2 second carrier PC2 represents the rotational speed of the device.	11	0	0	1	0	0	0
3	λ1は歯数比を示す。	λ1 is the gear ratio.	Λ1 tooth gear ratios is shown.	6	0	0	0	0	0	0
4	送風用のチャンバーは断面視して卵形形状としたチャンバー本体と、チャンバー本体の収れん部に設けたノズルとで構成した。	The chamber for ventilation includes a chamber body having an oval-shaped cross-section, and a nozzle provided in a collective part of the chamber body.	Chamber for ventilation of the oval-shaped cross section and a bottom chamber body 10, screws of a chamber body provided with a nozzle.	23	0	3	0	0	0	0
5	レンガ解体装置はその他の種々の構造のレンガ解体装置であっても良い。	The brick dismantling device may also be a brick dismantling device having various other structures.	The brick dismantling device brick dismantling device good even in a variety of other structures.	15	0	0	0	0	0	0
6	レンガ解体装置は先端に耐火レンガの解体部材が取り付けられた少なくとも一対の伸縮棒状部材からなるものである必要は必ずしもない。	The brick dismantling device does not necessarily need to be composed of at least a pair of telescopic rod-shaped members attached to a refractory brick dismantling members at the tip.	A distal end of the brick dismantling device fire brick dismantling member is attached to at least one pair of telescopic rod-shaped members is that which comprises all necessary.	29	0	0	2	1	0	1
7	NPC1は第一キャリアＰＣ１の回転速度を示す。	NPC1 represents the rotational speed of the primary carrier PC1.	NPC1 first carrier PC1 represents the rotational speed of the vehicle.	11	0	1	0	0	0	0
8	バイパス通路２０の両側には切換弁２８，３０を設ける。	A bypass passage 20 is provided on both sides of the change-over valves 28, 30.	A bypass passage 20 of the change-over valves 28, 30 on both sides is provided.	15	0	0	0	0	0	0
9	前記切り欠き部に前記電磁膨張弁の外方を覆うカバー部材を設けることもできる。	The above notched section can have a cover member to cover the outer side of the above solenoid expansion valve	The notch can be provided in the cover member for covering the outer side of the solenoid expansion valve	19	0	0	0	0	0	0

Post Editing Reduces Data Shortfall

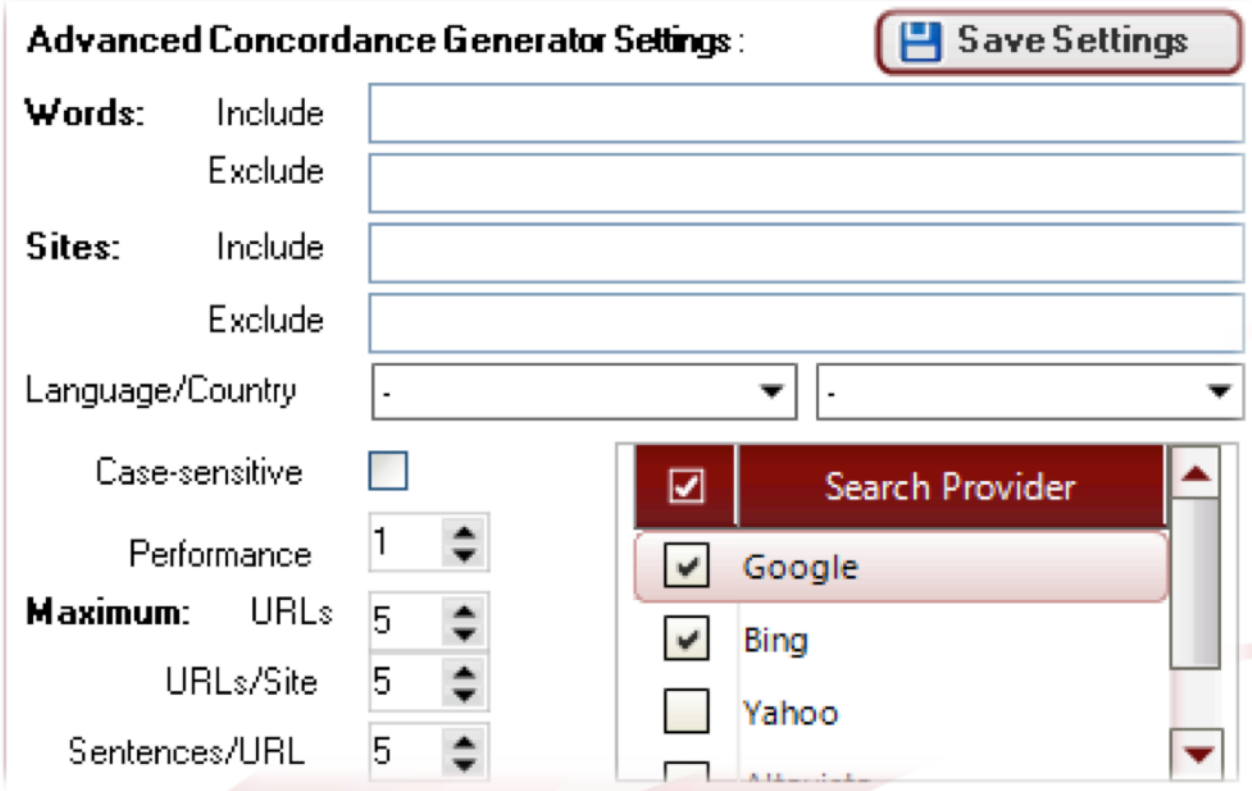


- Training data can often have gaps in coverage and an excess of data in other areas.
- Gaps in coverage reduce translation quality.
- Gaps can quickly be filled via post editing the machine translated output and submitting the data back to the system for further learning.
- Many gaps can be filled with monolingual data only.
- Further gaps can be identified and resolved by analyzing the text that is to be translated for high frequency terms and unknown words
- In some cases incorrect data may be statistically more relevant. Post editing will raise the relevance of the correct grammar.

Copyright © 2009, Asia Online Pte Ltd


Asia Online
The World Speaks One Language - Yours
Language Studio™ Pro
1.0.0.0
offline
Account : Asia Online
User : Demo
Menu
Corrective Data - Blunt's Soviet spying a mistake
Blunt's Soviet spying a mistake
Start Page
Corrective Data - Blunt's Soviet spying a mistake
Provide corrective data to improve future translations.
Overview
Monolingual Data
Bilingual Data
Phrase :
Add
Process
Stop
Phrase
Count
Dans cela 6
cela veut 5
Extracted from the corrections in the post editor
Advanced concordance search facilities
Scope All Select...
Find
Replace
Regular Expression
Whole Words Only
Match Case
Concordance
Audio Archive dans cela neant
whiteoutwhiteout - dans cela neant
whiteout" dans cela neant" epKuK 004december 2007free web-releaseThis audio is part of the collection: Open Source AudioArtist/Composer:whiteoutSource: http://www.koksundkohlede
Je n'ai aucun intérêt dans cela si ce n'est de rendre ou plutôt de donner l'occasion à des personnes de se rendre compte qu'ils peuvent gérer eux même leur santé ... et éventuellement aller plus loin en apprenant en autre les fondements de la Bio-Electronique Vincent qui est un système qui affranchi l'homme de ses peurs en lui apprenant que le milieu conditionne l'apparition de tel ou tel syndrome (ou maladie) et que de fait cet homme devient responsable de ce qui lui échoie, et donc ne peut se plaindre à personne des conséquences de son ignorance et des ses mauvaises habitudes de vie.
Il a plusieurs avantages uniques par rapport à d'autres systèmes dans cela agit en tant que passeport pour MSN, Office de phase, et emplacements de passeport de Microsoft, y compris les goûts de Hotmail (maintenant 5 gigaoctets de l'espace mémoire), espaces et OneCare.
aide dans cela je considère Vos suggestions et idées Seront fortement
Experts cela veut dire: "Misstinguette"?
French Language - Qu'est-ce que cela veut dire: "Misstinguette"?
qu'est ce que cela veut dire?
Contributions: ce que cela veut dire
5 commentaires pour "Contributions:ce que cela veut dire"
Advanced Concordance Generator Settings :
Save Settings
Words: Include
Exclude
Sites: Include
Exclude
Language/Country
Case-sensitive
Performance 1
Maximum: URLs 5
URLs/Site 5
Sentences/URL 5
Search Provider
Google
Bing
Yahoo
The best translations come from clean data. Tools determine clean data, but give user the chance to review and remove out of domain data and other unwanted data.
Save

Advanced Concordance Generation



Searches can be constrained and refined by many factors.

Example:

- **For Mexican Spanish, set language to Spanish and Country to Mexico.**
- **For Table (as in Database), set Sites to include "Oracle, Microsoft, ZDNet etc."**
- **To avoid overload of one particular style, limit the amount of return data by site.**
- **Add custom search providers such as existing TMS or public search engines.**

Post Editing Errors Amplifies Corrections

The quick brown fox over jumps the lazy dog

The quick brown fox jumps over the lazy dog

Additional corrective data generated by Language Studio™ Pro

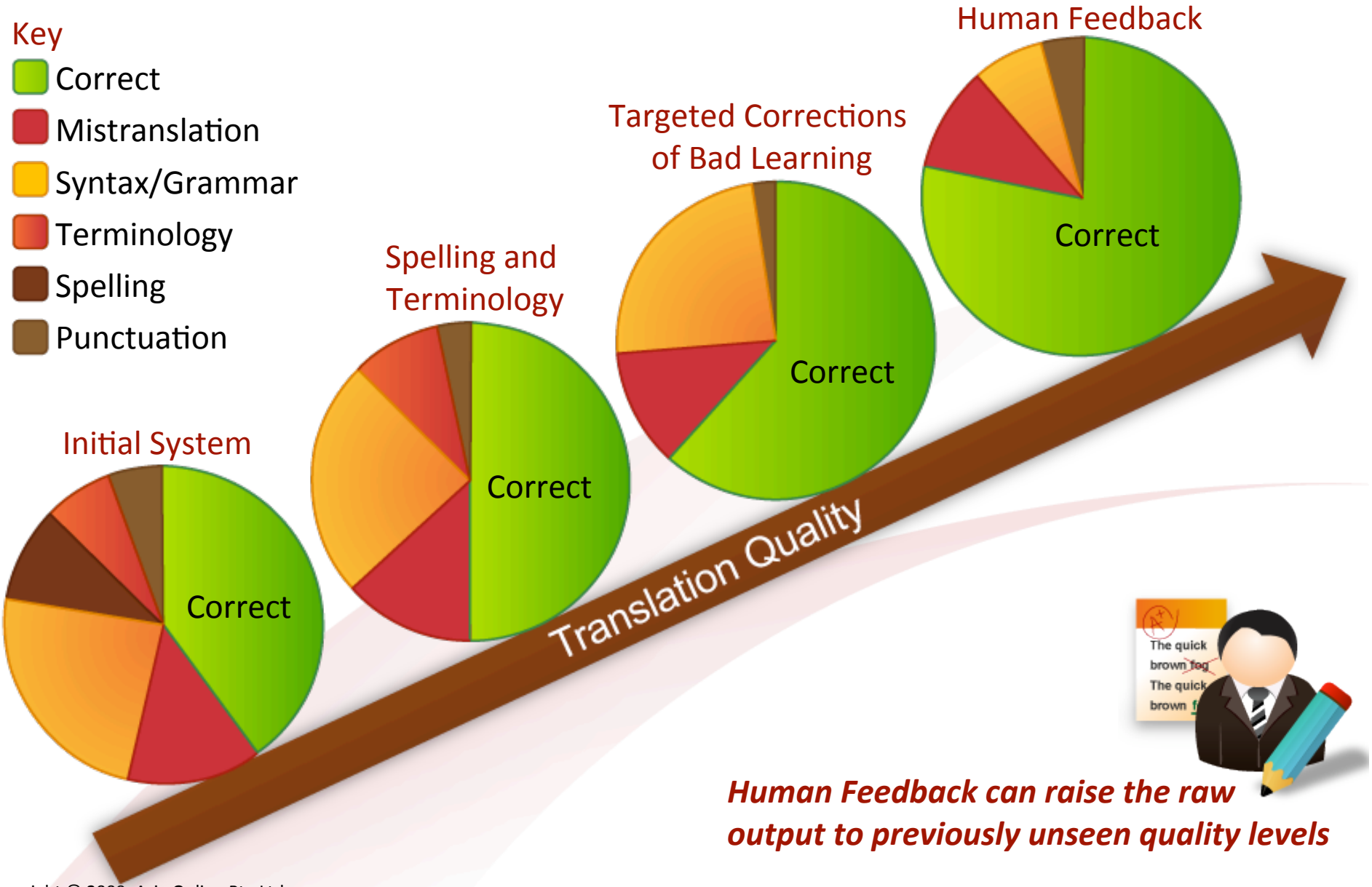
Buddha	**jumps over**	the wall
Siemens Wind Power CEO	**jumps over**	to Repower
Judge	**jumps over**	bench in courtroom melee
Military surveillance bot	**jumps over**	25 foot walls
Robbie Maddison	**jumps over**	Tower Bridge on motorbike
Man	**jumps over**	Grand Canyon
Cow	**jumps over**	Moon
With IE9 in sight, Firefox	**jumps over**	50% market share mark
Long jumper Brian Thomas	**jumps over**	a car to raise money
Rally car	**jumps over**	a crazy fan!
Kobe	**jumps over**	a speeding Aston Martin
A deer	**jumps over**	a motorcycle
A woman jogging in a California state park	**jumps over**	a 100-foot cliff to get away from attacker
An Afghan Army soldier	**jumps over**	a irrigation canal while conducting a foot patrol

Language Studio™ Pro analyzes corrections and generates other examples that include the corrected phrase to fill gaps in grammatical patterns. Each post edited correction is amplified producing many other corrective patterns, improving future translations.

Copyright © 2009, Asia Online Pte Ltd

Quality Evolution via Corrective Feedback

Asia Online™
The World Speaks One Language - Yours™



Human Feedback can raise the raw output to previously unseen quality levels

Copyright © 2009, Asia Online Pte Ltd

15 Domains – Focused Quality Translation

- The highest quality translation engines are specialized on domains.
- 15 pre-created domains provide a solid foundation upon which high quality customized engines for clients can quickly be created.
- A major element and accelerator when implementing stylized translations specific to clients.
- This is unique in the industry and gives Asia Online a significant competitive advantage.
- Domains can be combined.

	Domain
1	Law and Justice
2	Sports and Leisure
3	Technology and High-Tech
4	Military and Defense
5	Medical and Healthcare
6	Religion and Philosophy
7	Politics and Government
8	Economics and Finance
9	Science and Nature
10	Arts and Entertainment
11	Geography and Places
12	History and Events
13	Social and Culture
14	Blogs and Freeform Text
15	News and Commentary

Copyright © 2009, Asia Online Pte Ltd

Asia Online™
The World Speaks One Language - Yours™

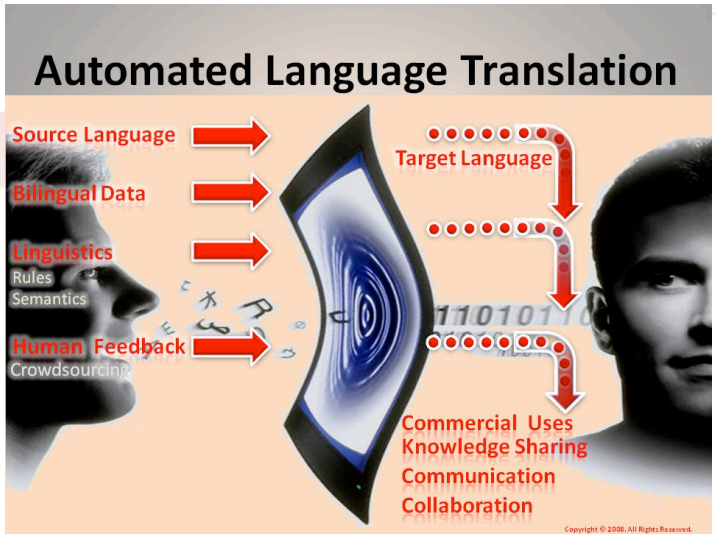
Thank You

kirti.vashee@asiaonline.net

Follow me on Twitter:
@kvashee

Join the Automated Language Translation Group in Linked In to monitor MT discussions

LinkedIn



Copyright © 2009, Asia Online Pte Ltd