SMART
Statistical Multilingual Analysis for Retrieval and Translation

SMART Final Review Meeting Introduction and Overview

Nicola Cancedda

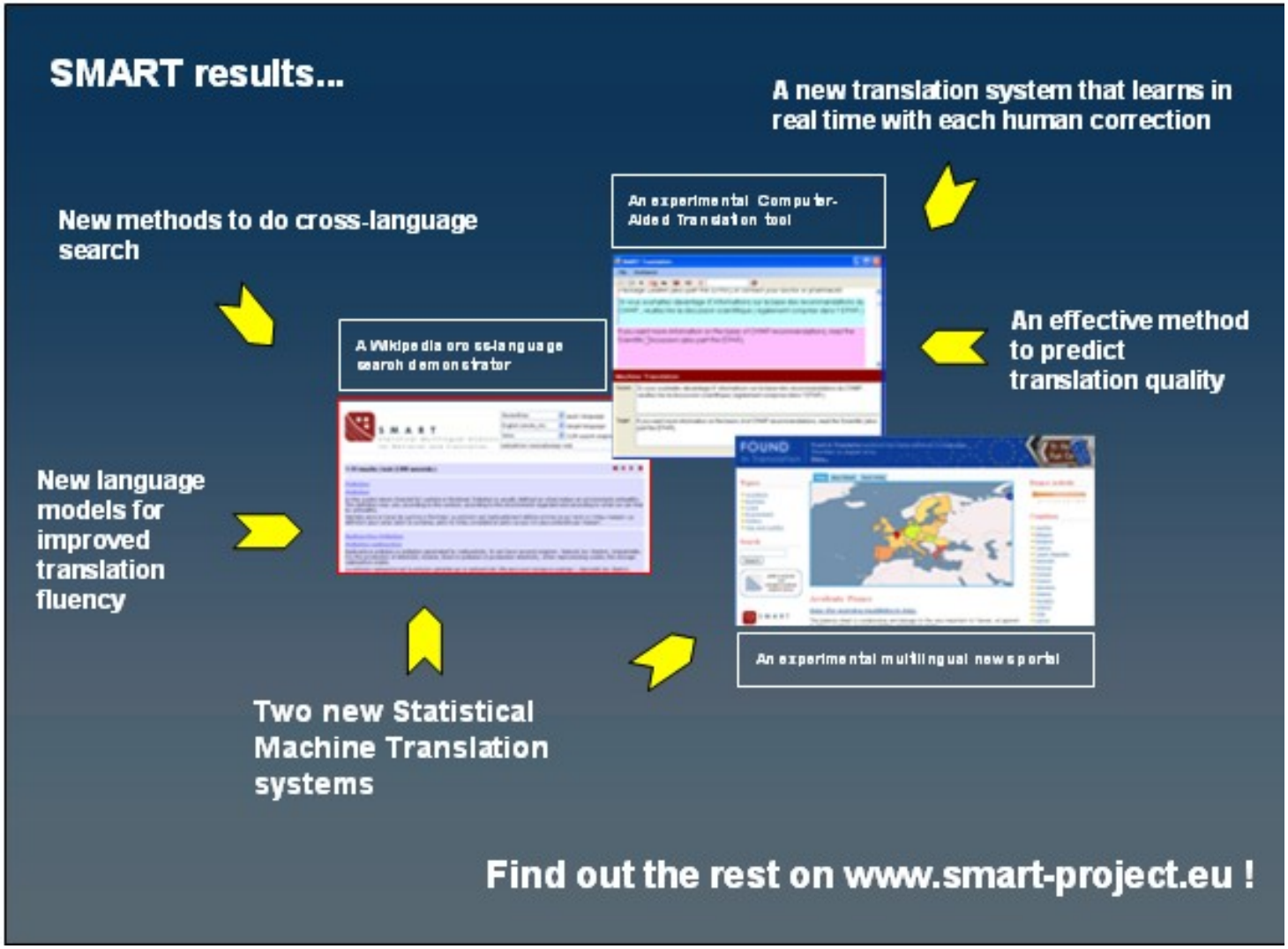
November 2009

SMART
Statistical Multilingual Analysis for Retrieval and Translation

- **Information Society Technologies Programme**
- **Sixth Framework Programme, "Specific Target Research Project" (STReP)**
- **Start date: October 1, 2006**
- **Duration: 3 years**
- **Budget: 3,526,379 €, of which 2,337,885 € of EC contribution**
- **~40 Researchers involved overall**

SIXTH FRAMEWORK PROGRAMME

Result Summary



The SMART Consortium

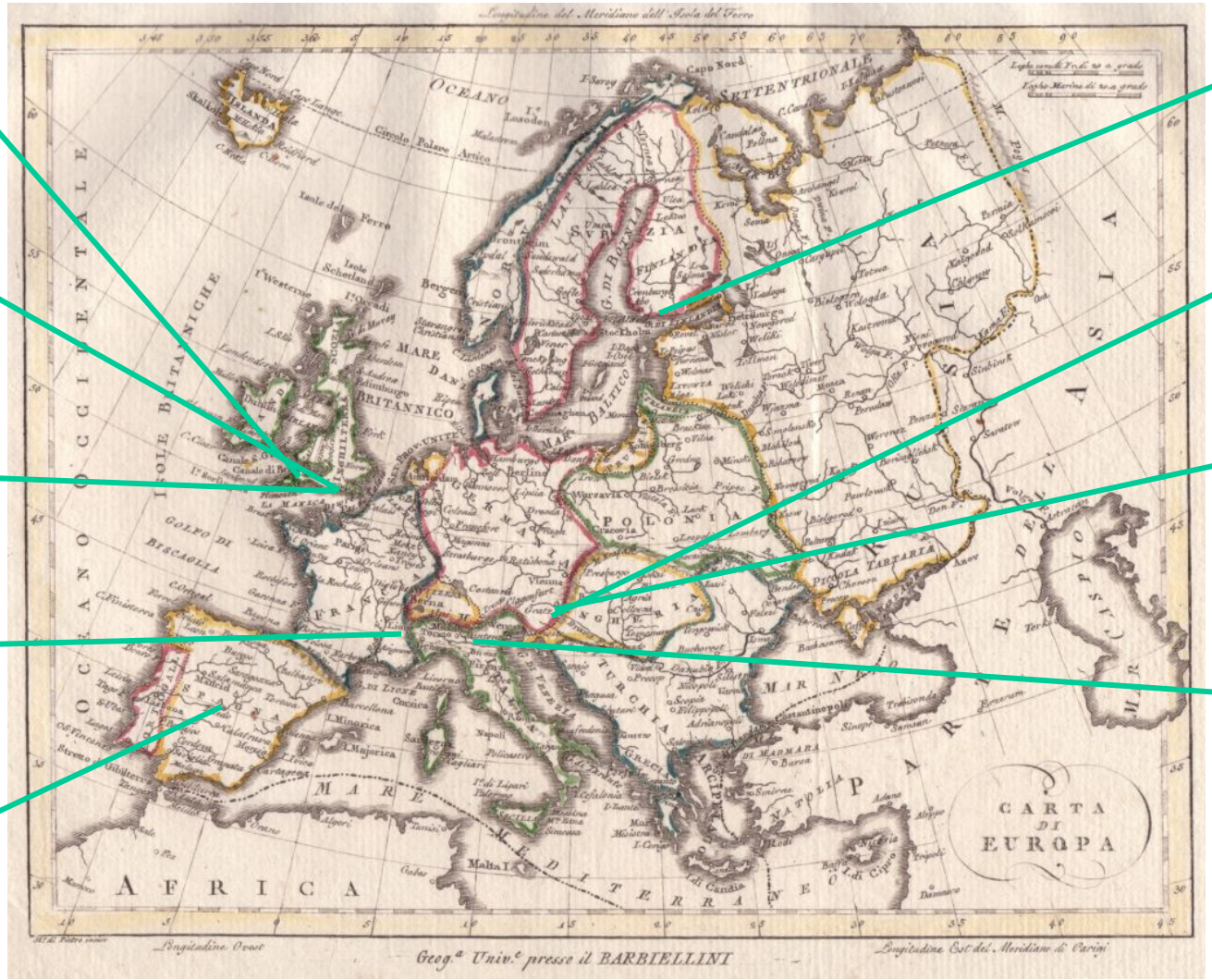
University of BRISTOL

UCL

University of Southampton

xerox

AMEBIS



The SMART Consortium

Motivation

From the proposal (2005)

- **Expanding global demand for tools for automatic translation and cross-language retrieval, clustering and categorization**
- **Statistical approaches largely successful, but still suffering from shortcomings, preventing their diffusion, e.g.:**
  - ***Relatively low fluency/grammaticality of output***
  - ***Model training still a somewhat arcane craft***
  - ***Difficult to use in new domains***
  - ***Trained once for all, no learning from constant user feedback***
- **Much incremental research, little risky disruptive research**
- **Some recent (in 2005) advances in Machine Learning very relevant for these tasks**

SMART is an attempt to bring Machine Learning closer to Machine Translation and Cross-Language Textual Information Access

Scientific and Technological Objectives

- **Problem: Mainstream two-layer approach to SMT results in tangled and opaque training:**
  - ***Proposal: Identify alternative formalizations for SMT amenable to exact solutions or approximations with performance guarantees***
- **Problem: SMT trained batch and keeps repeating the same mistakes**
  - ***Proposal: Design algorithms for on-line translation model training***

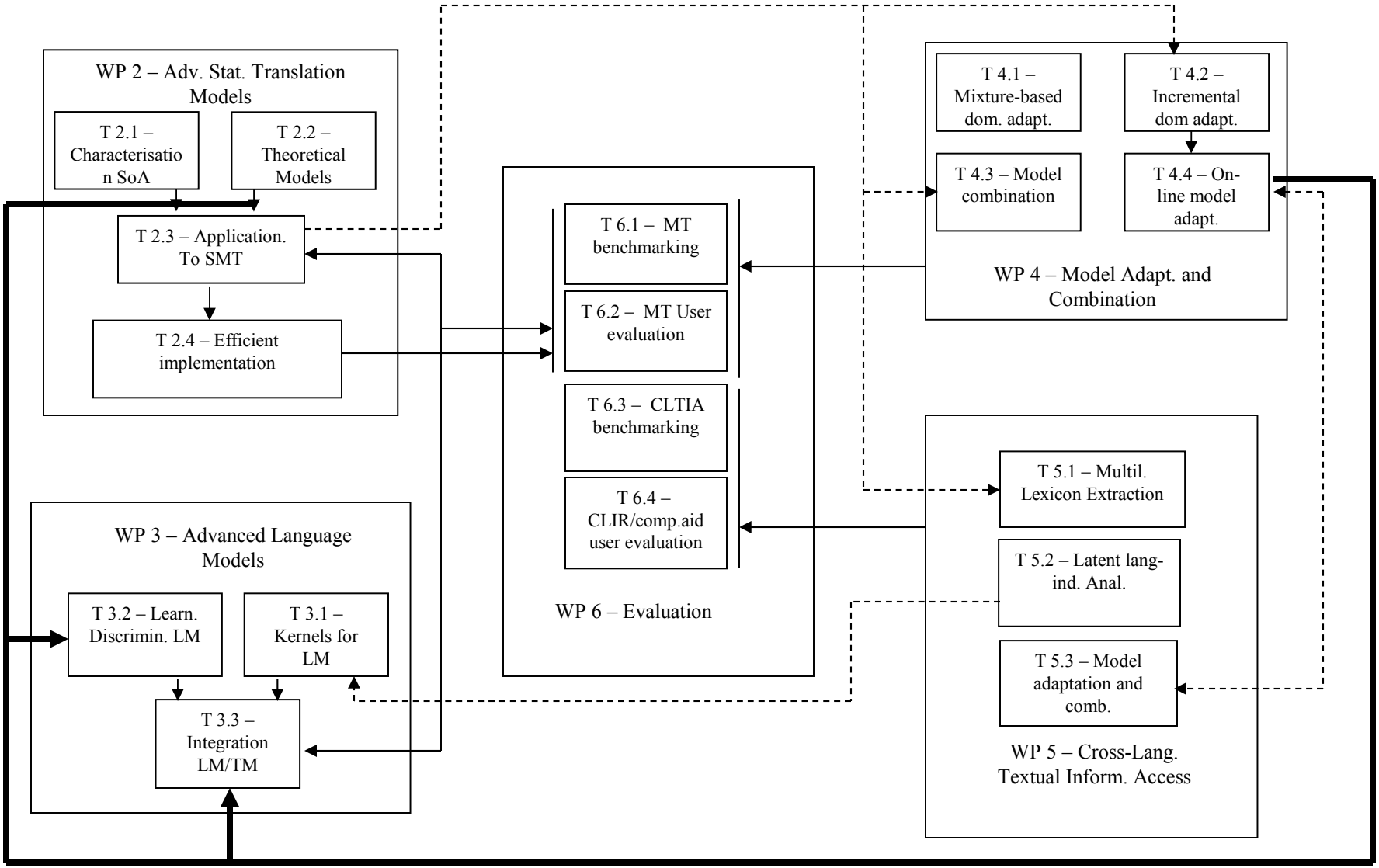
Scientific and Technological Objectives

- **Problem: SMT/CLTIA requires large amount of on-topic training data**
  - ***Proposal: Design models to combine large off-topic with small on-topic datasets***
- **Problem: Typical language models used in SMT are not trained directly to improve translations and are unable to capture morphological information**
  - ***Proposal: Discriminatively-trained language models benefiting from linguistic knowledge***

Scientific and Technological Objectives

- **Problem: Latent Semantic methods for CLTIA do not scale up to benefit from large datasets, or from three or more aligned languages**
  - ***Proposal: Investigate scalable approximations and multi-view variants; investigate synergies with phrase-based SMT***
- **Problem: Most CLTIA is limited to word-by-word translation and does not take context into account**
  - ***Proposal: investigate multi-word units for indexing and translation refinement based on context***
- **Problem: surface-form methods and latent-semantic methods are used alternatively**
  - ***Proposal: investigate what form of combination could do better than each one in isolation***

Workpackages



Needs & Skills

	Xerox	Amebis	Celer	JSI	NRC	UoB	UH	UniMi	USOU	UCL
Statistical Machine Translation	■				■					
Statistical Learning: Margin-based Methods						■	■	■	■	■
Statistical Learning: Kernels for structured data	■					■	■		■	■
Statistical Learning: On-line algorithms								■		
Computer-Aided Translation tools		■	■							
Call centre operations	■									
Cross-Language Information Retrieval	■			■			■			■
Cross-Language Categorization and Clustering	■			■			■		■	■
Multilingual Lexicon Extraction	■				■					
Creation and distribution of parallel corpora		■		■						

Expected Results

ER 2.1	**At least one incremental improvement leading to a global increment of X% in NIST score over the baseline under some training set size conditions. Baseline: NRC PORTAGE system as of March 2006** **X < 5: unsatisfactory; 5 ≤ X < 10: satisfactory; X ≥ 10: very satisfactory.**	
ER 2.2	**At least one totally innovative method coming within X% NIST score from baseline under some training set size conditions. Baseline: NRC PORTAGE system as of March 2006** **X > 20: unsatisfactory; 5 < X ≤ 20: satisfactory; X ≤ 5: very satisfactory**	
ER 3	**At least one Language Modeling method leading to a statistically significant improvement in the fluency of translations at an X% confidence level over baseline in a sentence-wise paired-t test. Baseline: 3gram model with Kneser-Ney smoothing (as implemented in the SRI LM toolkit).** **X < 95: unsatisfactory; 95 ≤ X < 99: satisfactory; X ≥ 99: very satisfactory**	

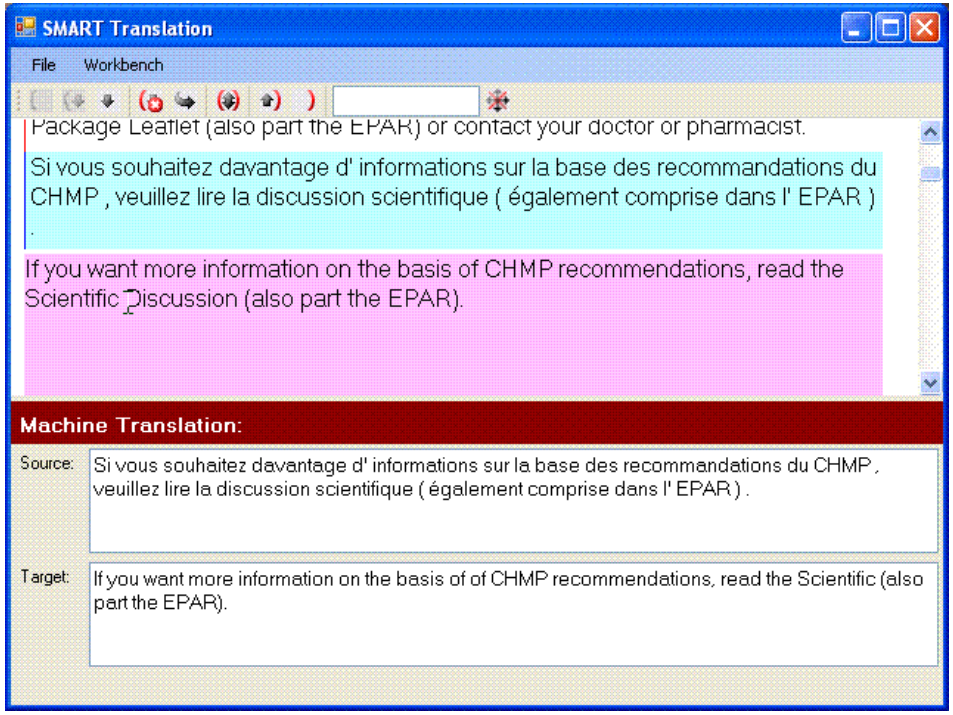
Expected Results

ER 4.1	**Method for adaptation when all (off-topic and on-topic) training data is available improving of at least X% NIST score over baseline under some training set size conditions. Baseline: PORTAGE system as of March 2006, trained on the union of off-topic and on-topic data.** **X < 5: unsatisfactory; 5 ≤ X < 10: satisfactory; X ≥ 10: very satisfactory.**	
ER 4.2	**Method for model adaptation when a seed model trained on off-topic data and an on-topic training dataset are available improving of at least X% in NIST score over baseline under some training set size conditions. Baseline: best score of PORTAGE system as of March 2006, trained on either the off-topic dataset used to train the seed model or the on-topic training dataset.** **X < 5: unsatisfactory; 5 ≤ X < 10: satisfactory; X ≥ 10: very satisfactory.**	
ER 4.3	**Method for combination of models previously trained on off-topic and on-topic training datasets respectively (actual training data unavailable) improving of at least X% in NIST score over baseline under some training set size conditions. Baseline: best score of PORTAGE system as of March 2006, trained on either the off-topic dataset used to train the seed model or the on-topic training dataset.** **X < 3: unsatisfactory; 3 ≤ X < 8: satisfactory; X ≥ 8: very satisfactory**	
ER 4.4	**Method for online adaptation of models improving of at least X% in NIST score over baseline under some training set size conditions. Baseline: PORTAGE system as of March 2006, trained on seed data only.** **X < 5: unsatisfactory; 5 ≤ X < 10: satisfactory; X ≥ 10: very satisfactory**	

Expected Results

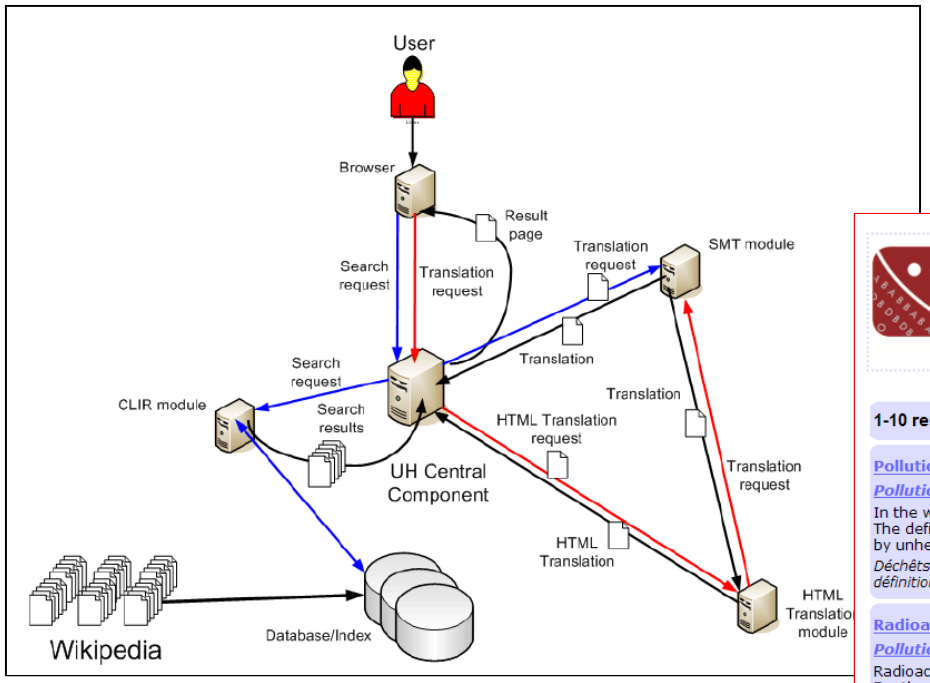
ER 5.1	**Method for CLIR significantly improving over baseline at an X% confidence level on Average Precision. Baseline: Xerox CLIR technology as of March 2006.** **X < 95: unsatisfactory; 95 ≤ X < 99: satisfactory; X ≥ 99: very satisfactory**	
ER 5.2	**Method for CLC significantly improving over baseline at an X% confidence level on F Score under some training set size conditions. Baseline: Xerox CLC technology as of March 2006.** **X < 95: unsatisfactory; 95 ≤ X < 99: satisfactory; X ≥ 99: very satisfactory**	
ER 6.1	**CAT: improvement in productivity of X% against baseline under some conditions of translation memory population. Baseline: same translation interface but no suggestions from the MT system.** **X < 3: unsatisfactory; 3 ≤ X < 15: satisfactory; X ≥ 15: very satisfactory**	
ER 6.2	**CLIR + comprehension aids, troubleshooters. Reduction of X% in average call time.** **X < 3: unsatisfactory; 3 ≤ X < 10: satisfactory; X ≥ 10: very satisfactory**	

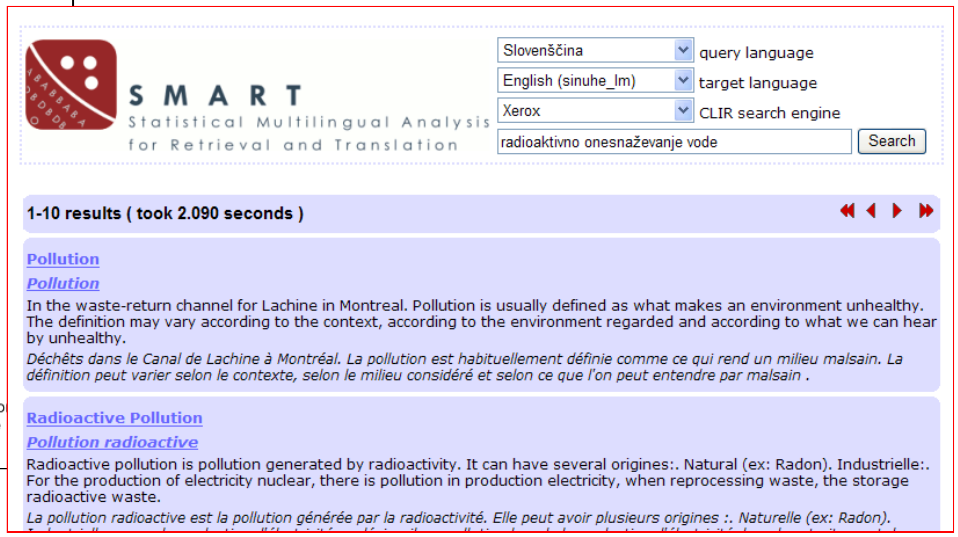
User Evaluations



- **Scenario 1: Computer-Aided Translation**
  - ***Can SMT effectively complement Translation Memory in increasing the productivity of professional translators? Does on-line adaptation help?***

User Evaluations





- **Scenario 2: Cross-Lingual Search on the Wikipedia**
  - ***Are casual users seeking to satisfy an information need using a document collection in a language they do not master more effective if they use CLIR technology to search and MT to translate results?***

User Evaluations

- **Scenario 3: CLIR and Comprehension Aids for Customer Service Representatives**
  - ***Are technical support agents in call centres more effective when accessing technical knowledge bases in English if they can use their own, non-English language for querying and can have technical terms translated in relevant pages?***

- **Replacement Scenario: Quality and Confidence Estimation**
  - ***Can the quality of automatic translation be accurately and confidently estimated without access to reference translations, and using only information normally available in CAT operations? Is this information useful to professional translators?***

Outline of the presentations (1/2)

10:00	WP 2, Advanced Statistical Translation Models, Sinuhe and MMBT (Craig Saunders)
10:25	WP 4, Model Adaptation and Combination, PORTAGE adaptive SMT (Nicolò Cesa-Bianchi)
10:40	WP 4, Model adaptation (Cyril Goutte)
10:55	Break
11:05	WP 3, Advanced Language Models, Xerox-UH re-ranking using Discriminative Language Models (Juho Rousu)
11:20	Confidence estimation (Nello Cristianini)
11:35	CAT demo + User Evaluation (Blaz Fortuna and Roberto Silva)

Outline of the presentations (2/2)

12:00	WP 5, Cross-Language Textual Information Access, Extensions of (K)CCA (John Shawe-Taylor)
12:15	WP 5, Lexicon adaptation and attempts at integration with CCA (Jean-Michel Renders)
12:30	Wikipedia demo + User Evaluation of the Wikipedia scenario (Kimmo Valtonen and Miro Romih)
12:55	Lunch break
13:55	Positioning SMART in the State of the Art (Cyril Goutte)
14:10	WP 7, Dissemination and Exploitation (Nello Cristianini)
14:25	Wrap-up and conclusions (Nicola Cancedda)