OpenMT: Open Source Machine Translation Using Hybrid Methods

- Kepa Sarasola, Iñaki Alegria
  Ixa Group. Euskal Herriko Unibertsitatea
- Núria Castell, Lluís Màrquez
  TALP, Universitat Politècnica de Catalunya
- Nerea Areta, Xabier Saralegi
  Elhuyar Fundazioa

TIN2006-15307-C03

http://ixa.si.ehu.es/openmt

Goals

- Defining open source architectures for machine translation.
- Combining different MT paradigms into a hybrid system.
  - Rule-Based MT (RBMT),
  - Statistical MT (SMT) and
  - Example-Based MT (EBMT).
- Integration of syntactic and semantic processing in machine translation.
- Evaluation.
  - A general open-source evaluation tool
  - Merging different criteria and using different languages and domains.

Goals by subprojects

- Coordinator: IXA (EHU)
  - Combining different MT paradigms into a hybrid system
  - Enrich it with semantics captured by ML from EBMT system.
- TALP (UPC)
  - Statistical MT technology needed for the combined MT systems
    - Enrichment with syntactic–semantic information captured by ML
    - Linguistically-based transfer rules from EBMT
  - Evaluation work package, adaptation of IQMT for MT
- Elhuyar
  - Basic infrastructure for MT resources and tools
  - Evaluation
  - Exploitation and Dissemination

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Basic tools and resources

- Web tool for managing and storing parallel corpora [51]
- Several new reference corpora
  - Collecting translation memories.
  - Domains: news, divulgative texts, manuals, environment and public administration.
- Elexbi: Extraction of specific domain bilingual terminology [2]
- Terminology extraction from comparable corpora [38, 39]
- Framework to represent tagged bitext [10,11]
- Semantic Role Labeling (ca, en & es) [41,34,33,42,40,32,44,45]

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Improving the actual RBMT and SMT systems

- Improvements in modules [3,4]
  - Lexicon, parsing (Freeling), translation of prepositions, ...
- en-eu language pair included.
- Comparing RBMT and SMT systems [28,43]:
  - Automatic metrics: RBMT < SMT on the in-domain data.
  - HTER evaluation : RBMT > SMT for both corpora.
- Improving basic SMT systems [21,20,18]:
  - Combining shallow-syntactic translation models based on linguistic data views
  - Domain adaptation issue by porting an en-es phrase-based SMT system

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Combining EBMT and RBMT

- Spanish-to-Basque **MultiEngine MT** [6,8]
- Specific domain (Labour agreements)

  Hierarchical strategy, selecting the best output:
  - first, EBMT translation patterns,
  - second, SMT (if its confidence score > a fixed threshold)
  - and, finally, RBMT.

  Important improvement in translation quality
  - 193.55% relative increase for BLEU (EBMT+SMT / SMT)
  - 15.08% relative increase (EBMT+SMT / EBMT)
  - Those improvements would be difficult to obtain by single-engine systems.

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Hybrid architectures based on SMT

- Two experiments on **Statistical Post Editing** (SPE) [12]
  - **Using a morphological component** in both RBMT and SMT
  - Limited size of the available corpora.
- Results are coherent with previous literature
  - **Large improvements using the RBMT+SPE on a restricted domain**.
  - 200% improvement on BLEU scores for a RBMT+SPE approach, when comparing to raw translation output
  - The relative improvement with respect to the single SMT system is 40%

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Advanced semantic processing for MT

- Inclusion of discriminative phrase selection models in SMT [23,24]
  - Local classifiers trained using WSD techniques instead of Maximum Likelihood.
  - Same techniques applied to English--Arabic, obtaining comparable improvements [13,14]
- Usage of external lexical-semantic resources to improve coverage and quality of SMT when applied across domains.
  - Enriching the original translation model with information derived from a multilingual LKB
  - Translation probabilities using a set of heuristics based on WordNet and local context.
  - Applied this to several translation tasks with significant improvement in translation quality.
- Inclusion of Semantic Role Labeling information (semantic dependencies) is left for the third year of research.

Project Progress and Achievements

- Basic tools and resources
- Improving actual RBMT and SMT systems
- Combining EBMT and RBMT
- Hybrid architectures
- Advanced semantic processing for MT
- Evaluation and demonstration

Evaluation of MT systems

- **Framework for MT automatic evaluation: IQMT** [19]
  - Set of heterogeneous metrics at different linguistic levels.
  - IQMT manual published [17].
  - software freely available: http://www.lsi.upc.edu/~nlp/IQMT/
- **Meta-evaluation of metrics**.
  - Combination of metrics and the assessment of the quality of metric sets.
  - Metrics based on syntax and semantics more reliable than the widespread lexical-based metrics as BLEU [22].
  - Considering also semantic features at the discourse level [18,25]
  - Study the behaviour of a non-parametric metric combination scheme [27]

Evaluation of MT systems

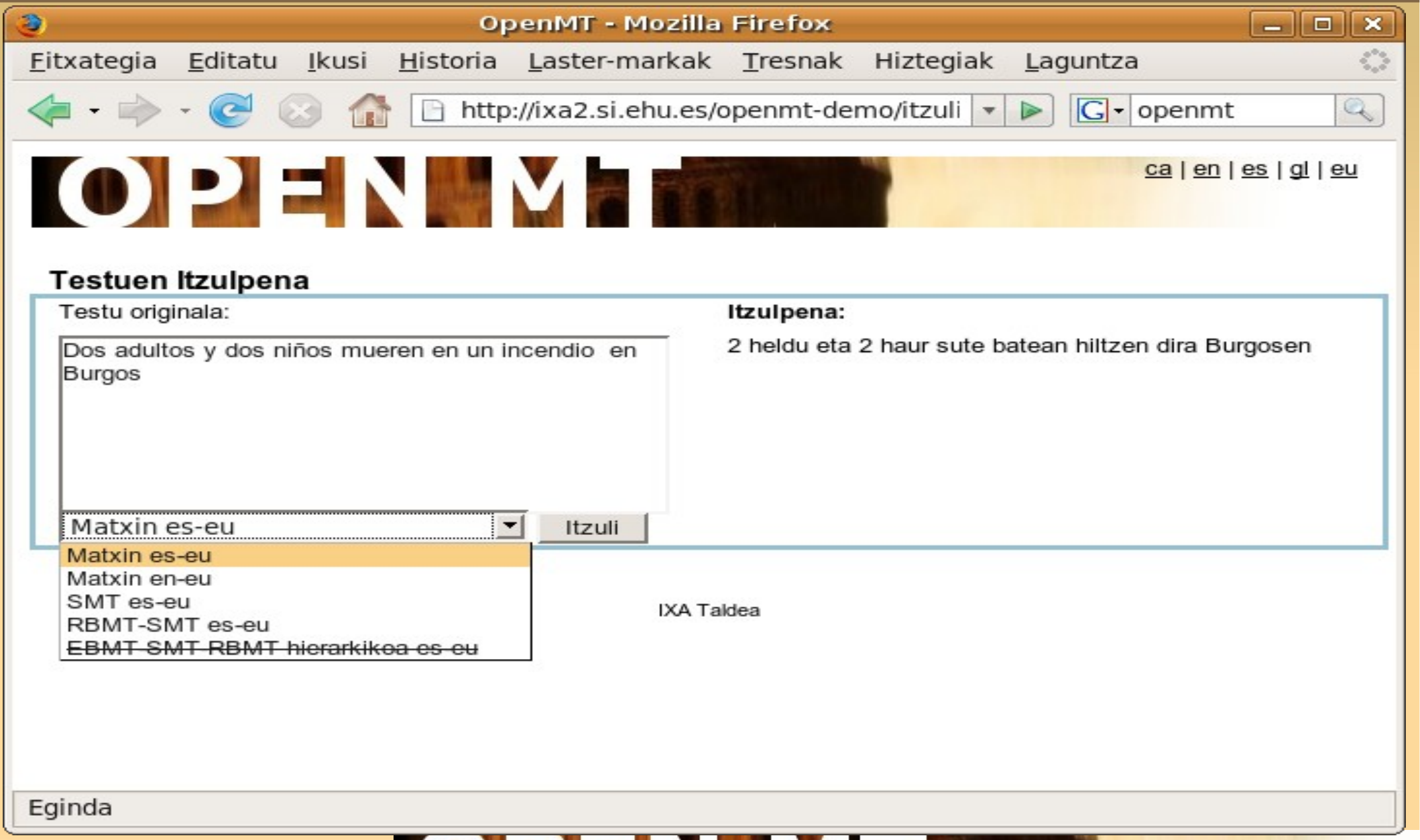
- The previous linguistic-based metrics and their combinations have been evaluated
  - ACL Workshop on Machine Translation 2007 [22]
  - ACL Workshop on Machine Translation 2008 [23]
  - NIST Metrics MATR workshop
- Other usages of the IQMT metric set explored:
  - Parameter adjustment for combinations with maximum descriptive power [30]
  - Viability of performing heterogeneous automatic MT error analyses [26]

Results and Indicators
IXA-EHU

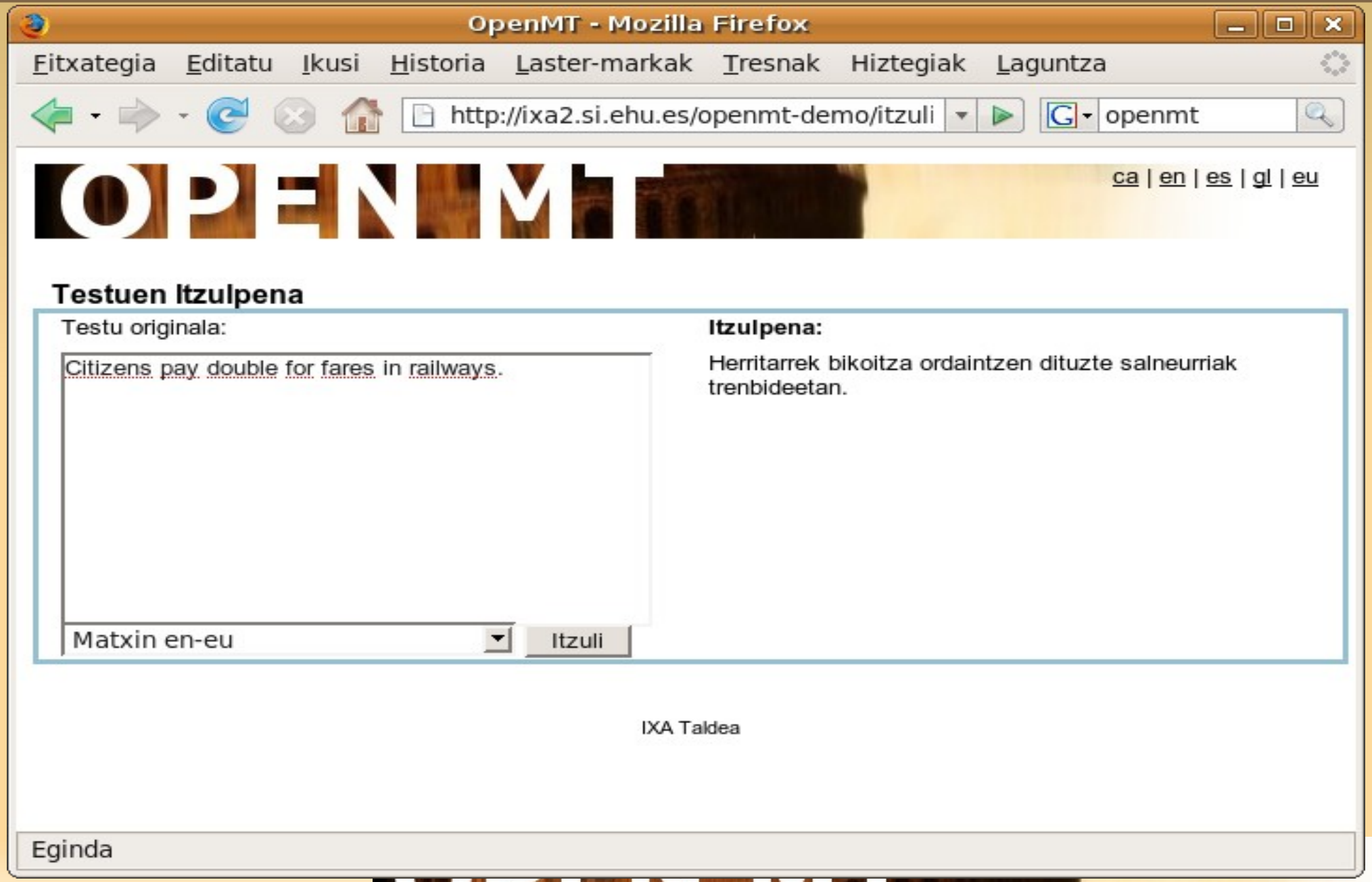
Results, relevance and production

- The goals have been achieved.
  - but RBMT for the en-eu pair postponed a few months.
  - Better results than initially expected in SMT and hybrid systems
- 15 publications and 2 demos.
  - Two papers in AMTA and one in MT-Summit, the main conferences in the MT area.
  - Publications in CICLING and in IJCNLP conference.

Demo:
MT from Spanish to Basque (4 systems)



Demo:
MT from English to Basque



Results and Indicators
IXA-EHU

Person training

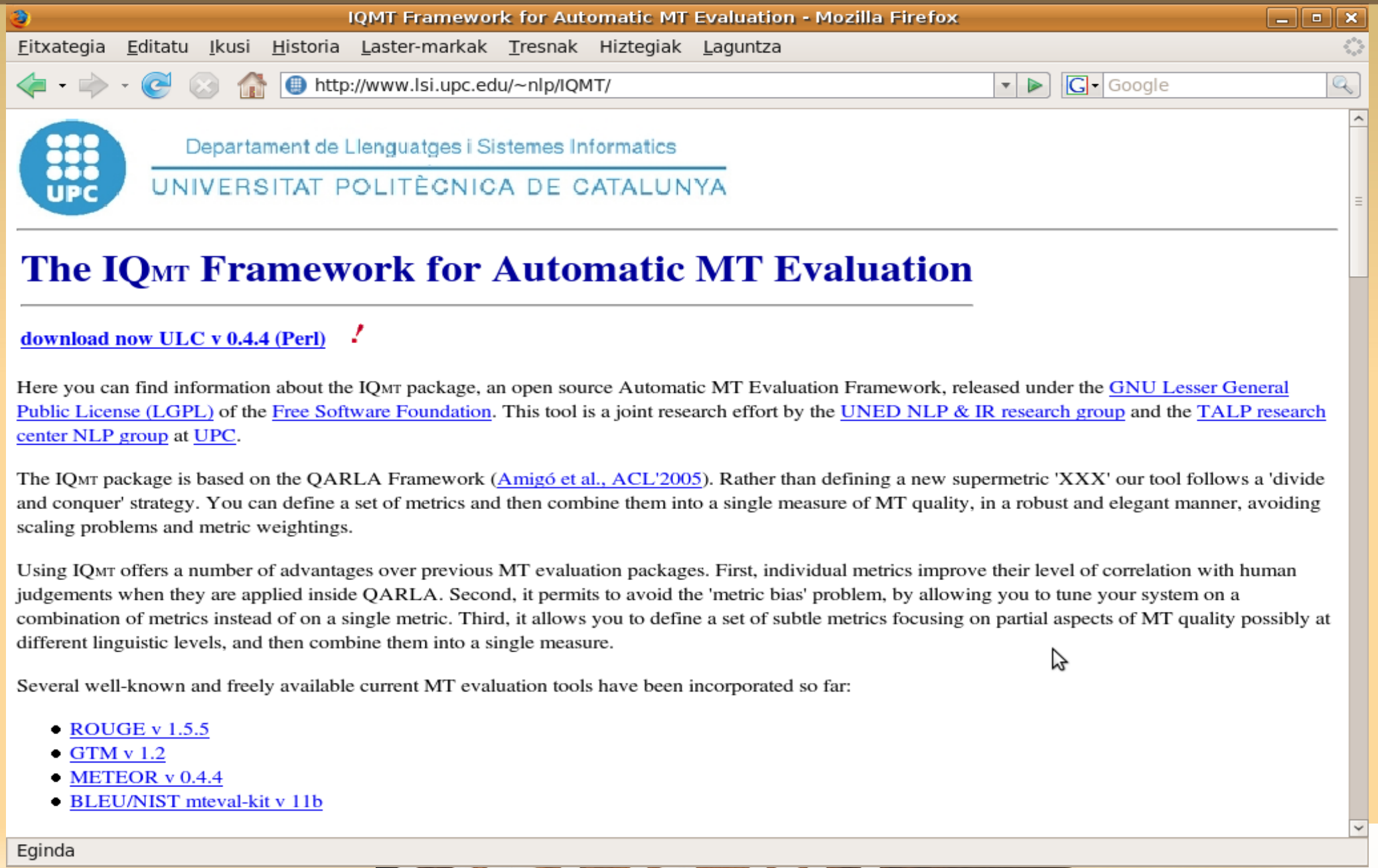
- Three students participate in OpenMT-Ixa.
  - One PhD thesis presented in 2007 by Aingeru Mayor [35]
  - Other PhD thesis to be presented in 2009 by Gorka Labaka
  - Gorka Labaka won the Albaycin evaluation award for MT integrated in the JTH2008 [29]
- Doctors from IXA are PhD advisors
  - 3 participants from Elhuyar: Antton Gurrutxaga (DEA), Xabier Saralegi and Igor Leturia. To be finished during 2010.
- Material from our project used in
  - HAP official master on LNP (ixa.si.ehu,es/master)
  - conferences and dissemination courses (UEU, www.ueu.org)

Results and Indicators
TALP-UPC

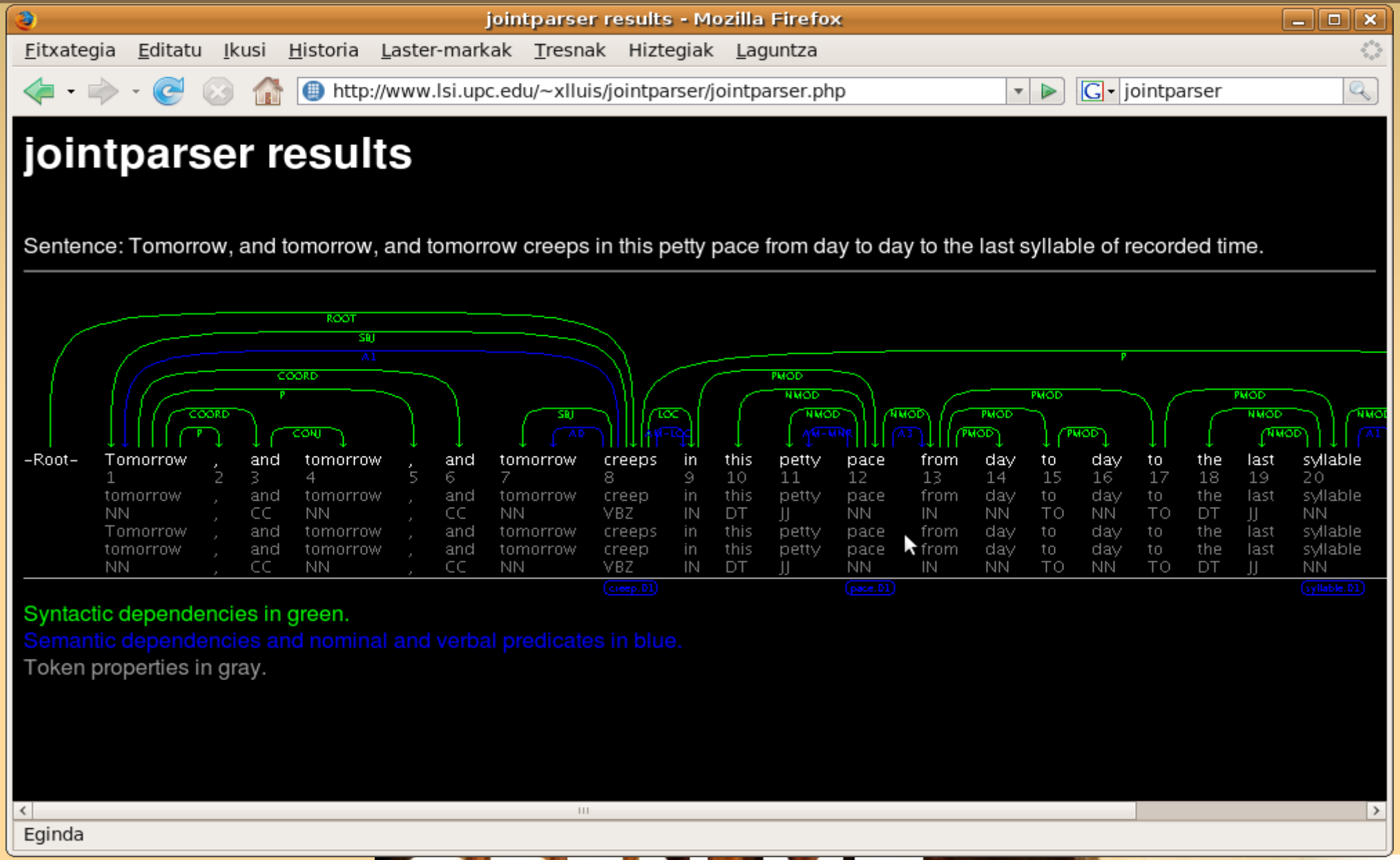
Results, relevance and production

- The goals have been achieved.
- 34 publications
  - 2 publications in Journal of Artificial Intelligence Research (JAIR). The impact factor of the journal is 1.107 (SCI/SSCI).
  - One publication in the ACL, the most prestigious in the NLP area
  - A 50-page chapter in a MIT Press book on Machine Learning for SMT (NIPS series)
  - Publications at the COLING-ACL and the IJCNLP conferences
- 2 demos:
  - JointParser joint analysis of syntactic and semantic dependencies. http://www.lsi.upc.edu/~xlluis/jointparser/jointparser.php
  - The IQMT Framework for Automatic MT Evaluation http://www.lsi.upc.edu/~nlp/IQMT/

IQMT
Framework for Automatic MT Evaluation



JointParser:
Joint syntactic and semantic analysis



Results and Indicators
TALP-UPC

Person training

- Three students participate in OpenMT.
  - PhD thesis presented in 2008 by Jesús Gimenez [18]
  - Master Thesis presented in 2008 by Cristina España [13]
  - Master Thesis presented in 2008 by Xabier Lluís [31]
  - Stefan Bott achieved a post-doc grant from the Catalan Government.

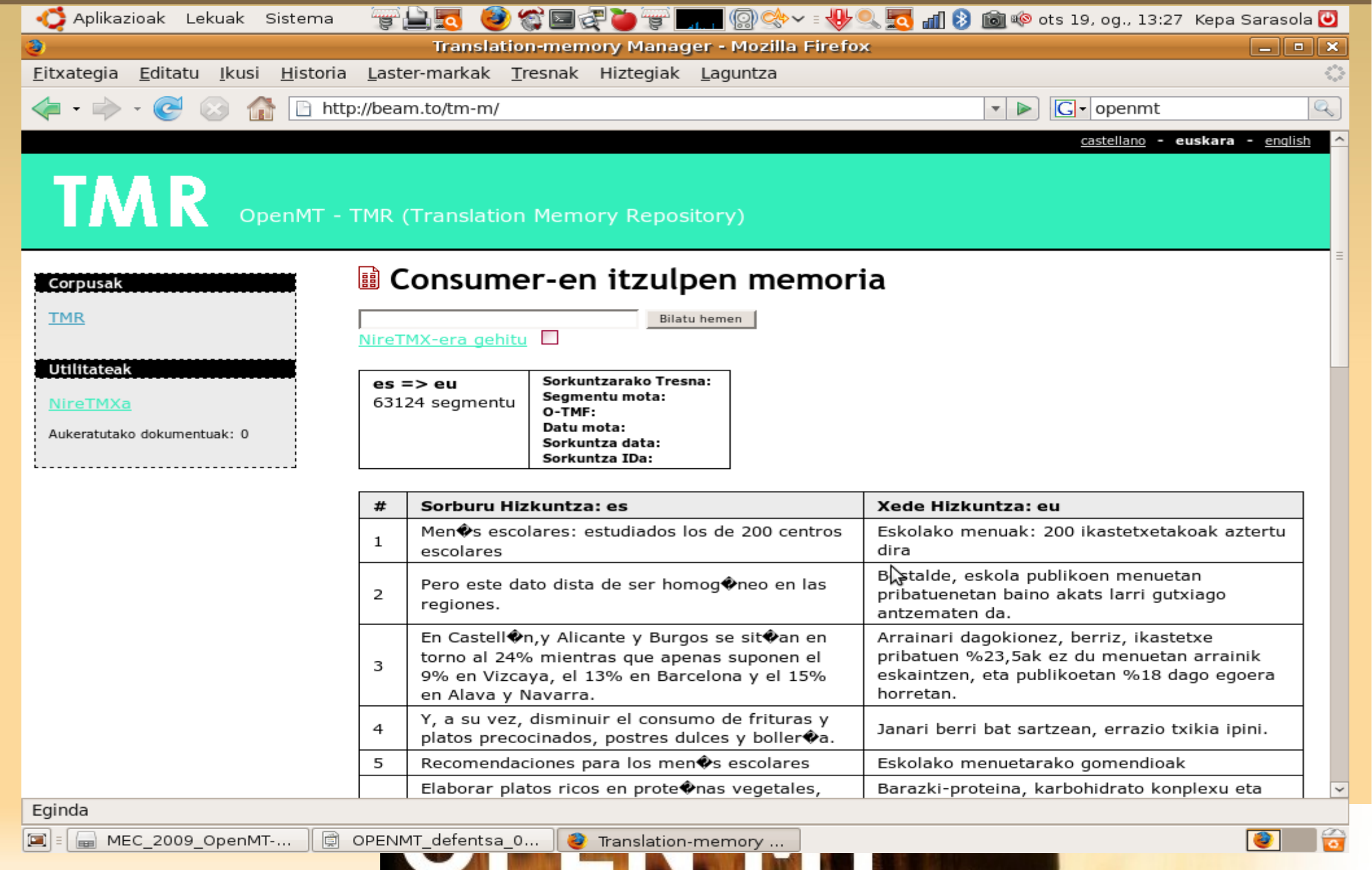
Results and Indicators Elhuyar

Results, relevance and production

- The goals have been achieved.
  - Parallel corpora collected for several domains
  - Tool for managing and storing parallel corpora
  - Bilingual lexicon from the administrative--text domain was created making use of the terminology extractor Elexbi.
  - Works on terminology extraction from comparable corpora
- 3 publications
- Demo: Website repository of Translation Memories

  http://ixa.si.ehu.es/openmt/demoak_html

Demo:
TMR translation Memory Repository



Results and Indicators Elhuyar

Person training

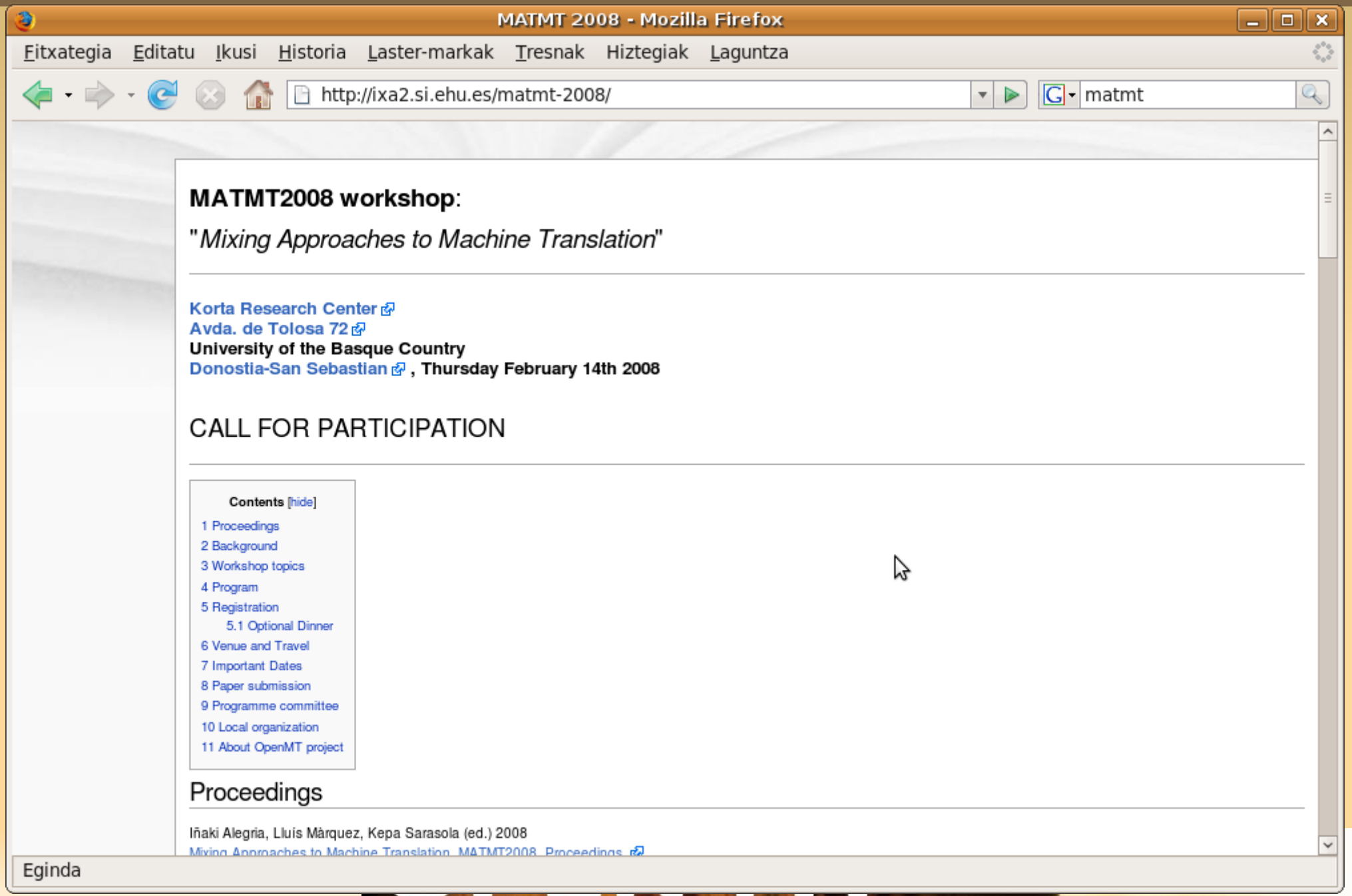
- Two participants finishing their Official Masters
  - Xabier Saralegi
  - Igor Leturia
- Another participant obtained the PhD research diploma (DEA) an is currently finishing the PhD thesis.
  - Antton Gurrutxaga

Results and Indicators
EHU + UPC + ELHUYAR

International collaboration (1)

- Workshop organization 2008/02/14:
  Mixing Approaches to Machine Translation [5,48]
  - More than 50 persons came from Spain, Europe and Japan.
  - Invited talks were given by M. Federico, P. Koehn and A. Way.
  - They were very useful for the main aim of the project: discussing different methods for hybridization in MT.
  - Dissemination of the results has been carried out
- Network of Excellence on MT designed during the MATMT workshop 2008 in San Sebastian.
  - Coordinator: H. Sommers from DCU university.
  - Consortium: Ixa group the DCU (Dublin), UPC (Barcelona), Edinburgh, Aachen and Charles (Praga) universities, Xerox company (Grenoble) and Bruno Kressler (Trento) Foundation.

OPENMT
MATMT workshop (2008/02/14)



Results and Indicators
EHU + UPC + ELHUYAR

International collaboration (2)

- STREP project was proposed in 2007
  - It was called EurOpenTrans (Large Scale Open Source Machine Translation for low density European Languages).
  - Partners included: DCU (Dublin), UPC(Barcelona), Edinburgh and Charles (Praga) universities, Alpinion (Ljubljana) and Translan (Dublin) companies and Elhuyar Foundation.
  - The project was evaluated with 9.0 points when the threshold was 10.0.
- Project proposal (TransBlog) in the European Work Programme 2009-2010, Challenge 2 call
  - Leader: Elhuyar. Partners: IXA-EHU.
  - The consortium is set up
  - Application of technologies developed in OpenMT in a framework for translating and increasing visibility of Blogs

Results and Indicators
EHU + UPC + ELHUYAR

International collaboration (3)

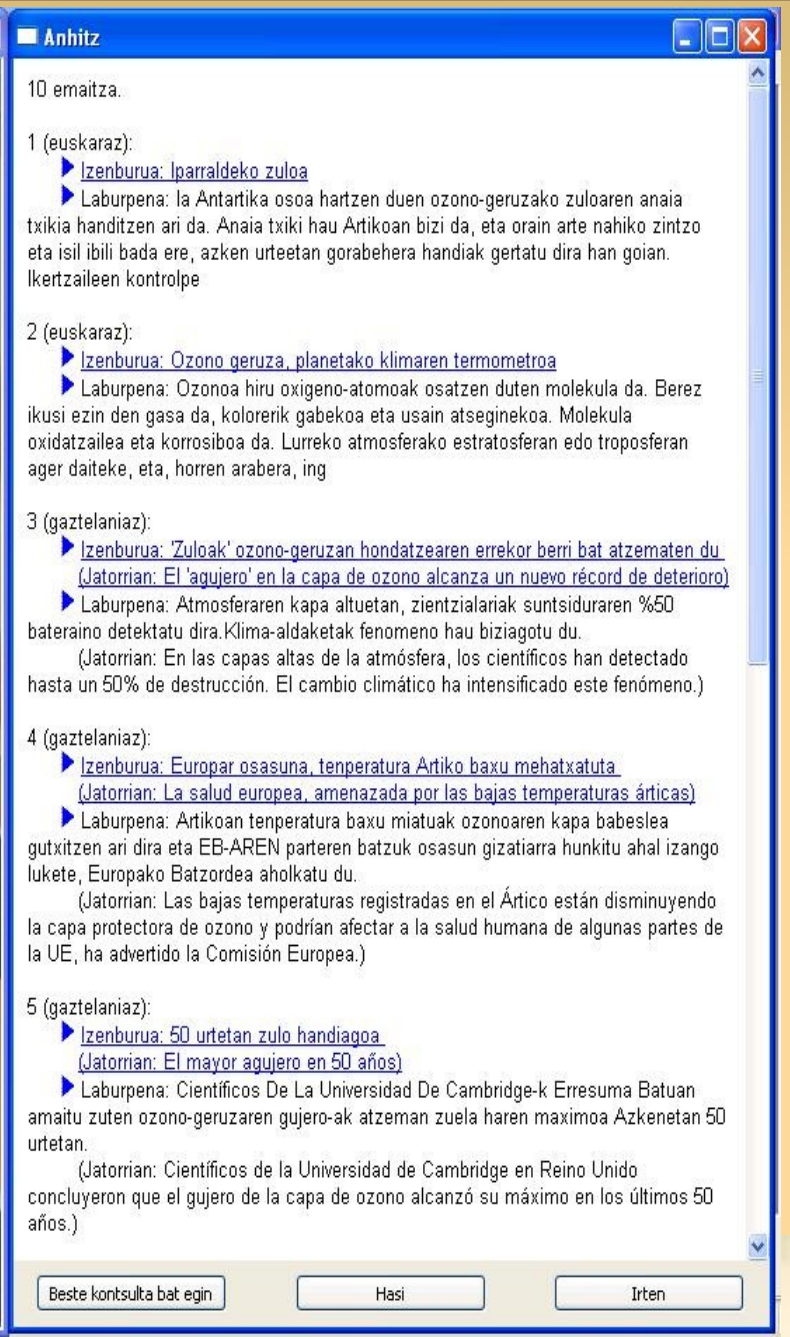
- Andy Way's group at the Dublin City University.
  - Kepa Sarasola (IXA-EHU): Stay of 2 months in 2006.
  - Gorka Labaka (IXA-EHU): 2 months in 2007+ 2 months in 2008
  - Two papers generated [43,28]

Results and Indicators
EHU + UPC + ELHUYAR

Technology transfer

- Most of the modules and technology are being transferred in the framework of two PROFIT projects to:
  - Eleka company (www.eleka.net)
  - OpenTrad consortium (www.opentrad.org)
- Presented to a public call for a major contract in es-eu MT Basque Government 2008.
- Integrated in a CLQA system named Ihardetsi [53] which was included in the QA-CLEF evaluation 2008.
- MT technology encapsulated in AnHitz project [52], promoted by the Basque Government.

ANHITZ:
Speech CLIR and QA with avatar and MT

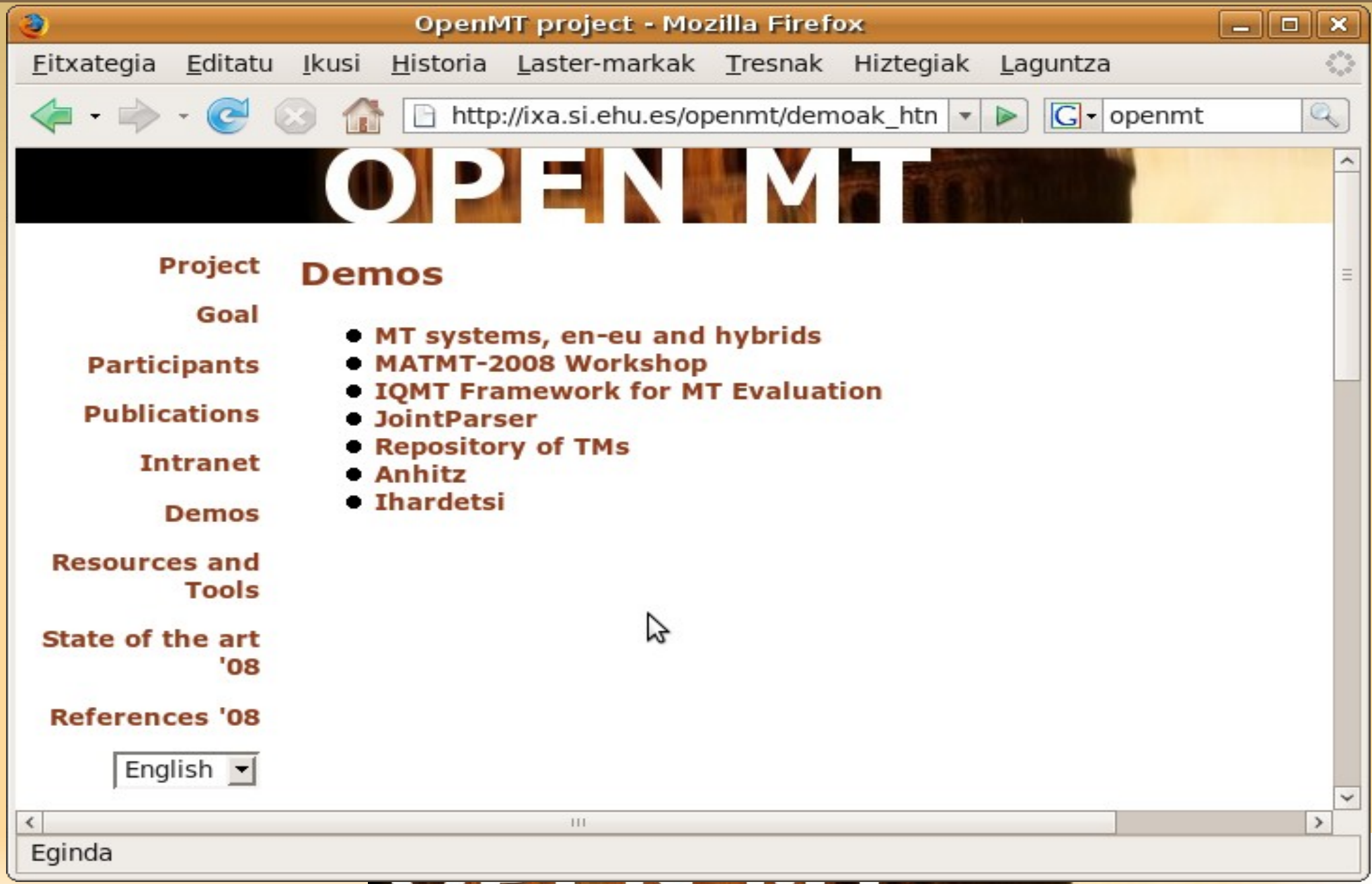


Results and Indicators
EHU + UPC + ELHUYAR

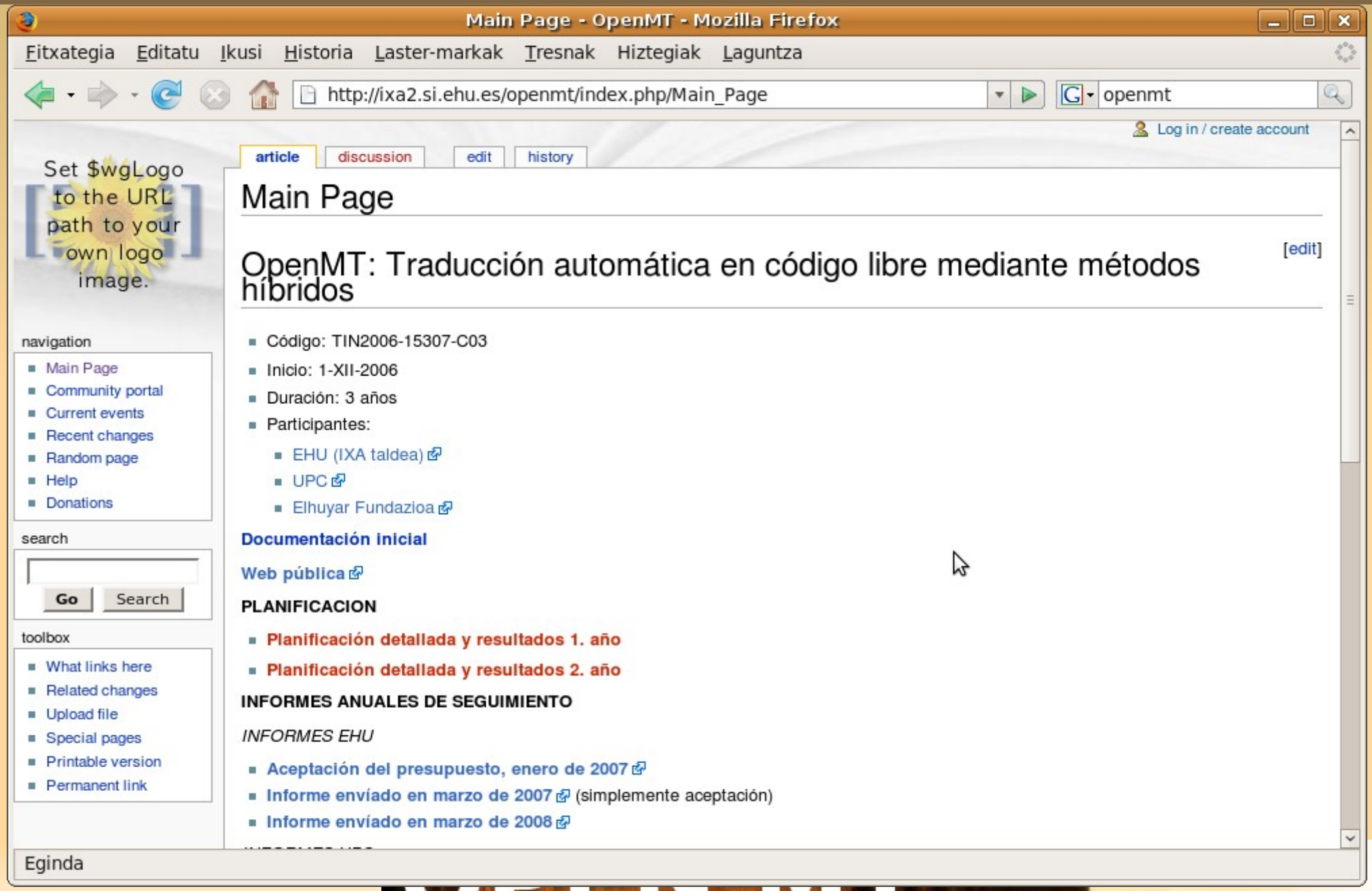
Project management

- Private web page of the project (wiki): http://ixa2.si.ehu.es/openmt
- Public website of the project: http://ixa.si.ehu.es/openmt
- Periodical meetings
  - (1) Donostia 19/20-XII-2006,
  - (2) Barcelona 9-X-2007,
  - (3) Donostia 14-II-2008,
  - (4) Barcelona: 2-X-2008.
- Technical meetings:
  - (1) Lexical selection, Donostia 19-IV-2007;
  - (2) Lexical selection, Barcelona 6-VII-2007,
  - (3) MT-hybridization, Donostia 14-II-2008.

OPENMT
public web site



OPENMT
private web site



OpenMT: Open Source Machine Translation Using Hybrid Methods

- Kepa Sarasola, Iñaki Alegria
  Ixa Group. Euskal Herriko Unibertsitatea
- Núria Castell, Lluís Màrquez
  TALP, Universitat Politècnica de Catalunya
- Nerea Areta, Xabier Saralegi
  Elhuyar Fundazioa

TIN2006-15307-C03

http://ixa.si.ehu.es/openmt