Infoday,
Bxl 26 Jan 09

Roberto Cencioni

Multilingual Web

Theme 5 of the ICT-PSP Workprogramme 2009

DG Information Society and Media

Unit INFSO.E1
Language Technologies
& Machine Translation

infso-e1@ec.europa.eu

European Commission
Information Society and Media

Background

- **the Web revolution**
  - – the Web gives unlimited access to a wealth of information
  - – online collaboration, social networking … enable Web users to become content producers & remixers

  … but significant **language barriers** remain
- **the enlarging Europe with 23+ languages**
  - – we don’t have enough translators!
- **single European information space**
  - – one of the main objectives of the i2010 policy framework:
    - ▪ bridging language barriers in the InfoSoc

European Commission
Information Society and Media

Basic facts

- **EU official languages: 23 x 22 = 506 pairs**
  - EC MT (Systran core engine) has 18 pairs in operation & 10 more pairs at prototype stage
  - 60+ regional & minority languages within the EU
- **English accounts for 30% of today's Web content**
  - 50% in 2000, 35% in 2004
  - Arabic, Chinese, Portuguese … growing very fast
- **nearly 1,5 billion internet users worldwide** (2008)
  - c 320 million native EN speakers in the world
- **key requirements for the "digital translation market":**
  - volume
  - access
  - personalisation
    - real quick, real cheap

“Europe’s language is Translation”

- **translation & interpretation market**
  - c $15 billion worldwide (40% in Europe); €1.1 billion for EU institutions alone (2006)
  - est. 300,000 full time salaried translators worldwide (37% in Europe)
- **a good European base**
  - SDL, Star, RWS, XRX, Euroscript, Logos, Moravia, VistaTEC, Semantix …
  - ESTeam, Lucy Software …
- **a largely untapped potential**
  - 4x according to some companies

Business world

- **new models:** Collaborative, web-based technologies allow translation to become more agile, faster, and better with fewer steps (CSA Inc.)

- **new markets:** Language Weaver is entering the three new strategic markets – Web Content, Business Intelligence and Customer Care – to provide high-volume, high-speed, and accurate automated translation solutions

- **new approaches:** If you don't see your native language here, you can help Google create it by becoming a volunteer translator. Check out our Google in Your Language program.

European Commission
Information Society and Media

In a nutshell

support & enhance

- interpersonal & business **communication**
- **information** access & publishing

– for everybody
– **across languages**
– emphasis on online environments

- research & technology: **ICT Call 3**
- services & validation: **ICT-PSP Call 4**

ICT-PSP Call 3,

Approach

- **innovative & effective combination of people, processes & technology**; the end result is not science, rather
  - more and/or better output
  - save time
  - cut cost
- **solution oriented:** useful & useable although possibly not perfect, think ROI
- user/industry driven; time horizon: 3-5 years
- based upon robust although possibly commercially untried technology
- convincing **use scenario** & target domain, real(istic) data volumes & flows
- emphasis on **evaluation**
  - adequate plans, resources & metrics
- credible risk analysis & **exploitation channels** …

Scope

- **Machine Translation (MT)** as defined in the work-programme encompasses
  1. fully automatic machine translation
  2. interactive computer-aided translation (eg TM)
  3. a suitable combination of 1. and/or 2. with web based
     - human translation, proof-reading & post-editing aids including where relevant methods inspired from social networks
     - content management & workflow systems …
- emphasis on **language transfer**, from source language to target language(s)
  - language **input-output** (e.g. speech-to-text) is not the focus
  - cross-platform, multi-format content **access/delivery** is key

Language coverage

- some of the work is expected to be language independent
  - flexibility & ease of adaptation to other languages are key factors!
  - content authoring & management, collaboration & workflow … are language independent anyway
- project outcomes should be validated in 3+ languages
  - preferably belonging to different linguistic families
  - with the aim of broadening language coverage wrt. existing commercial offerings
- languages are chosen & justified by the proposers bearing in mind the following priorities (from high to low):
  1. **EU official languages**
  2. regional languages
  3. minority languages
- Non-EU world languages linked to global markets & exports can be considered as well
  - on a proposal by proposal basis

Expectations

- **impact is key**, so: viability, sustainability, exploitation channels, deployment prospects …
- main findings must be pro-actively disseminated; some form of **public showcase** is mandatory
- **participants** should include
  - private or public sector content owners & aggregators
  - service providers & technology suppliers
  - providers of language services
  - (online) communities of interest where relevant
  - research centres where justified …
- 4-7 partners/project, up to €2.5 million funding, up to 36 months

European Commission
Information Society and Media

ICT-PSP Call 3,

Theme 5: Multilingual Web

5.1 machine translation for the multilingual Web (projects)

- information access: MT and other multilingual solutions for information access & analysis, esp. cross-lingual search & retrieval
- information publishing: MT to create, distribute and (re-)use more widely & effectively online content in a multilingual environment

5.3 multilingual Web content management (projects)

- communication: multilingual Web content development & management; design, authoring, versioning & maintenance of multilingual Web sites, portals or repositories

5.2 best practices & standards for the multilingual Web (network)

ICT-PSP, 5.3
Multilingual Web content management

- methods, techniques, metrics … for developing & managing multilingual web content & services
  - more than translation; **significant <u>cultural</u> elements**
- think of
  - one big website in many languages, or
  - several interrelated websites, one country/language each
- now think of how to maintain the integrity & consistency of such resources, effectively & over a long period of time
  - and how to detect & repair gaps or inconsistencies
- so, **beyond the "translation" step** (obj 5.1):
  - design, authoring, versioning & maintenance of (multiple, parallel, interconnected …) websites, portals or repositories
  - in a distributed collaborative environment, possibly across organisational boundaries
  - so as to turn a multi-million endeavour into a viable proposition for a much broader range of companies & administrations

European Commission
Information Society and Media

5.1 can be seen to some extent as a subset of 5.3
(its “translation box”)

- different **usages**:
  - web at large, enterprise, public information repositories …
- different **users**:
  - teams as well as individuals, engineers as well as analysts, sales & marketing, language professionals, … you & me
- different content rich, information based **sectors**, private & public
- **quality** depends on task & user
  - from raw translation & “gisting” up to error-free translation
- **two important conditions**:
  - widely recognised, well argued **problem;** clearly identified **user base**; credible **exploitation prospects**
  - thorough **validation** in a given domain / for a given task
    - volume
    - metrics

ICT-PSP, 5.2
Standards & best practices

Thematic network

- covers the **same broad issues as 5.3**
  - "the web as THE vehicle for multilingual content & services"
- provides a **forum for multilateral exchange of experience & consensus building**
- structure & tasks to be defined by the proposers, indicative list:
  - bring together a meaningful subset of the main stakeholders, possibly through their own groups & associations
    - ICT & language industries, content aggregators/distributors, e-services, multinational agencies, industry & de-jure standards bodies …
  - analyse current situation, identify gaps & bottlenecks; assess market failures if any, specify technical & non-technical conditions to be met and the respective actors
    - establish roadmap (trends, requirements, dependencies …) for further developments in the coming years
  - stimulate consensus & active involvement/coordination; take part in leading conferences, liaise with primary associations etc.
    - explore means to promote best practice (conferences, portals, publications, training …) beyond current channels
  - identify & describe suitable follow-on actions

What we don’t want

Not supported under this Call:

- proposals that do not address « language transfer »
  - yes: focus on mapping a source language into one or several target languages
- developments addressing immediate commercial concerns
  - no: simply adding a language pair to an existing product
- proposals requiring extensive R&D efforts
  - no: proposals centred around ‘unfinished’ research prototypes
- approaches that do not promise to deliver performance along with portability, scalability & maintenability
  - yes: emphasis on automation, flexibility & cost effectiveness
  - no: labour intensive coding of linguistic knowledge

Practical info

ICT-PSP Theme 5 – Multilingual Web

budget: 14 Meuro under Call 3

managed by: Unit E1

EC contact: Mr Kimmo Rossi

email: infso-e1@ec.europa.eu

- inquiries: from the call publication date
- pre-proposals: until 3 weeks before the call closing date

search for experts: online registration form at

http://ec.europa.eu/information_society/activities/ict_psp/cf/expert/login/index.cfm

FP7-ICT Call 4
at a glance

a) **Core research exploring new avenues for MT**

- ground breaking, multidisciplinary, high risk – high promise research
- architectures & technologies that learn and adapt flexibly & effectively to different languages, domains & tasks
- catering for new forms of language & communication (eg online communities; dynamic, volatile …)

b) **Problem oriented research for specific tasks & usage contexts**

- online translation for the masses
- translation in distributed collaborative environments
- managing multilingual communication & content
- automatic acquisition & annotation of language resources

c) **Community building & networking**

- reinvigorate European machine translation (MT) community
- build bridges between MT & MLT and other relevant disciplines
- help develop & coordinate shared technical infrastructure, promote reusability & interoperability, foster evaluation

European Commission
Information Society and Media

Thank you!

cordis.europa.eu/fp7/ict/language-technologies/..

FP7-ICT: ../**fp7-call4_en.html**

ICT-PSP: ../**cip-psp_en.html**