

Human Language Technologies @ EC

Roberto Cencioni

European Commission

INFSO - Information Society & Media

Digital Content & Cognitive Systems



Mission statement

- teach computers how to understand & process **written & spoken human language**
 - information
 - communication
 - interaction
- if you master language, then you can try & cope with **multiple languages**
 - nickname: **HLT** – several terms, communities & specialist groups:
 - natural language processing
 - speech technology
 - machine translation
 - information extraction
 - ...

A long-term commitment

- **EC has supported HLT** since 1970s:
 - sustained R&D effort throughout the 1990s
 - pioneering MT & TM technologies
 - relatively low-level profile in recent years
- **a fresh start** since 2008:
 - renewed political commitment, after the enlargement
 - explosion of online content, in languages other than EN
 - promising S&T advances, mostly linked to data-driven approaches

Scale of the challenge

- **EU has 23 official working languages**
 - 60+ languages in Europe
- **English accounts for a mere 29% of Internet content**
 - BRIC & other languages growing much faster
 - English native speakers account for 27% of Internet users
- **eCommerce: 2/3 of EU customers only buy in their own language**
- **Europe accounts for 50% of the worldwide language services market**
 - and yet users & professionals cannot cope with huge & volatile volumes of web content

“Europe is still a patchwork of national online markets, and Europeans are prevented from enjoying the benefits of a **digital single market**. Commercial and cultural content and services need to flow across borders.”

Challenges, internal

- **the sector can do better** in terms of
 - **credibility** = useable results & uptake
 - **critical mass** = clear directions & shared agenda
 - **visibility** = public & political awareness
- players must **address fragmentation**
 - link research communities & specialist groups, academia & research labs, vendors & leading users
 - pool, share, reuse basic methods, tools & datasets
 - enhance result-oriented cross-border collaboration
- ... **before FP8 starts**, within 3 years

EU financial instruments

- **current programmes**
 - research & **technology** (FP7 **ICT**)
 - competitiveness & **innovation** (CIP **ICT-PSP**)
- dedicated investment in the HLT area:
 - 2008 0
 - 2009 40 M
 - 2010-11 ~83 M (est.)
 - 2012-13: ?
- ~55 projects by mid-2012

Innovation programme (PSP, 2009-10)

– *emphasis on SMEs & less-resourced languages*

- **pilot projects = demonstration**
 - demonstrate the potential of existing technology
 - emphasis on service innovation in real(istic) settings
- **LR actions = infrastructure**
 - assemble LRs, improve their (re)usability, make them available in open repositories
 - emphasis on organisational build-up & sustainability
- focused efforts: 30 M over 2 years

State of play upcoming...

Research programme (ICT, 2011-12)

- 12 calls over 2 years
- largest calls scheduled for Sept 2010 & July 2011
- HLT part of **Challenge 4** "*Technologies for Digital Content & Languages*"
- appears in **2 calls**:
 - **Call 7**: open Sept, close Jan 2011, 50 M
 - **SME call**: open Feb 2011, close Sept 2011 (2-stage process), 35 M

4.2 Language Technologies

- **basic elements:**
 - both **written & spoken language**
 - **multilingual** (in/out), where relevant cross-lingual
 - handle **everyday language**
 - cope with **massive volumes** & diverse sources
 - **contextualisation & personalisation**
 - technologies are **adaptive** (language, domain, task)
 - but... **embedding & testing** within specific (demanding) application environments

Objective 4.2 overview

- **3 research lines (“outcomes”)**
 - multilingual content processing
 - information access & mining
 - natural spoken interaction
- **no predefined budget allocation**
- **balanced mix of projects**
 - 50% STREP (21 M)
 - 30% IP (13 M)
 - 20% open (8 M)

Objective 4.2

research lines

a. multilingual content processing

- addresses the **production** chain in a multilingual setting (authoring, translating & publishing)
 - exploit language-encoded knowledge embedded in documents, social media, web & audio-visual objects
- two project lines:
 - **advance machine translation** on several fronts
 - quality, self-learning & adaptation...
 - everyday language, x-lingual resources...
 - **test & improve suitability** (usability, effectiveness...) of novel technologies in real-life settings
- instruments: IP (1) + STREP

Objective 4.2

research lines

b. information access & mining

- **finding, interpreting, correlating, categorizing...**
digital content
 - exploit language-encoded knowledge embedded in documents, social media, web & audio-visual objects
 - combine linguistic, statistical, semantic... approaches
- progress towards **broad coverage** coupled with (efficient) **deep analysis**, in multiple languages
- in one or several of the following domains:
 - **cross-lingual information retrieval**
 - **audio & video mining**
 - **text mining**, from multilingual sources
- instruments: STREP

Objective 4.2

research lines

c. natural spoken interaction

- progress towards richer, more spontaneous & robust **man-machine** interaction
- **“conversational social agents”** that can
 - handle conversational speech, in & out
 - cater for social cues, in & out
 - learn from interaction, react to new situations...
- technologies that are
 - portable, non-intrusive, real-time...
- either **component technologies** or complete **proof-of-concept systems**, within larger ICT systems
- instruments: IP (1) + STREP

Objective 4.2

cross-cutting actions

d. coordination & support

- **unifying vision** & compelling **technology roadmap** for the field at large
- closer collaboration with **industry**, better understanding of the **demand** side, more active **user** involvement
- flexible, coordinated **evaluation** framework
- enhance fitness, (re)usability, interoperability of language data & tools by means of **pooling, trading & sharing**
 - virtual: **standards** i.e. methods, guides, best practices...
 - virtual & physical: **open repositories** of research results, development/training resources...
- instruments: CSA

4.1 SME initiative

- data is the crude oil of today's R&D and yet often too expensive for new or small actors
- ease development & first-use experimentation of novel technologies by **high-tech SMEs**
 - by **pooling & reusing** datasets & related tools
 - language data, see obj 4.2
 - knowledge (linked) data, see obj 4.4
- 3 intertwined dimensions for language players
 - fast, effective data **acquisition & aggregation**
 - digital **trading places**, open exchanges or commons
 - (experimental evidence of) **new or better services** resulting from combining, extending, repurposing... resources
- instruments: STREP (26 M) + CSA (9 M)

Objective 4.1 overview

- **budget:** 35 M for two communities (Know, Lng)
- **publication:** 1st Feb, 2011
- **2-step submission & evaluation:**
 - short synopsis (5 pages), by 28 Apr
 - if successful, full proposal (50 pages), by 28 Sept
- **compact consortia:**
 - up to ~6 private/public partners
 - at least 2 SMEs, 30% of overall EU funding
- **focused projects:**
 - up to 24 months, up to 2 Meuro funding

Objective 4.1 definitions

- an SME is an **enterprise** which has
 - fewer than 250 **employees**
 - an annual **turnover** not exceeding 50 M
 - or an annual **balance-sheet total** not exceeding 43 M
- **relationships** with other enterprises must be taken into account
- the **official definition** of SMEs as per 2003/361/EC can be found at

http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index_en.htm

Key dates (tbc)

- **27-29 Sep 2010:**
 - ICT conference in Brussels, launch of Call 7
 - E1 ready to handle inquiries & pre-proposals
- **mid-Nov 2010**
 - dedicated HLT session(s): Lux 11/11 + Bxl 17/11
- **mid-Jan 2011:**
 - close of Call 7 - Language Technologies
- **Feb 2011:**
 - launch of SME-DCL call
- **close of SME call:**
 - Apr 2011 (1st stage, short proposals)
 - Sep 2011 (2nd stage)

Critical mass

- **quality is key**
- but **quantity matters** – and helps to stimulate competition thus preserving quality
- **3 focused calls in 2009-10**
 - 70 submissions
- **2 broader calls in 2010-11**
 - ~100 submissions?
- **academics to bring vendors & users!**

Thank you!

info-e1@ec.europa.eu

Upcoming ICT-HLT events (under construction):

http://cordis.europa.eu/fp7/ict/language-technologies/upcoming_en.html