Language Technologies @ EC, State of play

Roberto Cencioni

European Commission

INFSO - Information Society & Media

Digital Content & Cognitive Systems

Oslo, 18 June 2010

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Mission statement

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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Scale of the challenge

- 60+ languages in Europe
- **EU has 23 official working languages**
- **English accounts for 29% of Internet content**
  - BRIC & other languages growing much faster
- English native speakers account for 27% of Internet users
- **Europe accounts for 50% of the worldwide language services market** (mainly translation & localisation)
- and yet users & professionals cannot cope with huge & volatile volumes of Web 2.0 content
- **eCommerce: 2/3 of EU customers only buy in their own language**

"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."

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Challenges, internal

- **the sector can do better** in terms of
  - **credibility** = useable results & uptake
  - **critical mass** = clear directions & shared agenda
  - **visibility** = public & political awareness
- you 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

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Challenges, external

- **adapt to the socio-economic environment**
  - economic austerity
  - strong competition between policy areas
  - insufficient political drive for diversity
  - "multilingualism is expensive"
- **HLT is and will be challenged**
  - is it any good? does is work?
  - can't we leave to Google?

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

INFSO drivers

- **To become a recognized player within the <u>Digital Agenda for Europe</u>,** released 19 May 2010

  http://ec.europa.eu/information_society/digital-agenda/index_en.htm

  - **policy, (co-)regulation, benchmarking, …**
  - **research & innovation**
    - technological leadership
    - economic growth & jobs
    - societal challenges
  - **in a number of intertwined domains**
    - single digital market
    - public online services, eGovernment
    - digital skills & inclusion
    - trust & security, online safety …

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

EU financial instruments

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

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

State of play
ongoing…

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 Meuro over 2 years

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

State of play
upcoming…

Research programme (ICT, 2010-11)

- **workprogramme disclosed in July**
  - 5 calls end July, 12 in total over 2 years
  - largest calls scheduled for Sept 2010 & July 2011
  - **HLT part of Challenge 4 "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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

How about LRs?

- automated and/or collaborative **compilation of x-lingual LRs** e.g. from the web & digital collections
  - first series of projects underway; more under upcoming call
- "hub" services for **sharing & reusing** LRs:
  - META-SHARE (underway); further support for data exchanges under next call
- SME-driven **pooling** of LRs:
  - special call in early 2011
- creation, annotation… of **domain/task specific** LRs:
  - within the relevant technology-driven project

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

How about META-NET?

- **3 years (2010-12), 3 axes**
  - MT research, new x-disciplinary avenues
  - language resources
  - shared vision & roadmap for HLT at large
- **LR hub (META-SHARE)**
  - a multilateral exchange facility, not a research infrastructure
  - aimed at researchers, developers & professionals
  - stress on ICT technologies, services & applications
  - "keep it simple, make it happen", must pay for itself, hence
    - no heavy infrastructure, no long-term curation, no guaranteed QoS …
- initial **software & services** provided by META-NET partners
- initial **resources** supplied by META-NET & ongoing/upcoming INFSO projects

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Promoting the field

- **community services** launched within the next 18-24 months:
  - **unifying vision & technology roadmap** for the field at large (META-NET + others)
  - closer collaboration with **industry**, better understanding of the **demand** side, more active **user** involvement (new)
  - enhance fitness, (re)usability, interoperability… of LRs by means of **trading, pooling & sharing** (META-SHARE + others)
  - flexible, coordinated **evaluation** framework(s) (new)

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

ICT 2011-12
overview

- *HLT home within the ICT WP:*
  "**4.2 Language Technologies**", Call 7, 50 Meuro
- **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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
4.2 research lines

- **simple model**
  - "conventional" themes likely to foster new partnerships across disciplines, languages…
  - ambitious goals, with useful shorter-term spinoffs
- **balanced mix of projects**
  - 50% STREP (21 M)
  - 30% IP (13 M)
  - 20% open (8 M)
- **3 research lines ("outcomes")**
  - (multilingual) content processing
  - information access & mining
  - natural spoken interaction
- **no predefined budget allocation**

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
4.2 research lines

a. multilingual content processing

- get to and exploit (language-encoded) knowledge embedded in documents, social media, web objects
- for the purposes of **authoring, translating & publishing** online digital content
- two project lines:
  - **advance machine translation** on several fronts
    - quality, self-learning & adaptation…
    - everyday language, x-lingual resources…
  - **test & improve suitability** (usability, performance, effectiveness…) of novel technologies in real-life multilingual settings
- instruments: IP + STR

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
4.2 research lines

b. information access & mining

- get to and exploit (language-encoded) knowledge…
  - same as in a.
- for the purposes of **finding, interpreting, correlating, categorizing**… online digital content
- progress towards **broad coverage** coupled with (efficient) **deep analysis**, in multiple languages
- apply to one or several of the following:
  - **cross-lingual information retrieval**
  - **audio & video mining**
  - **text mining** from multilingual sources
- instruments: STR

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
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 + STR

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
4.2 cross-cutting

d. coordination & support actions

- overcome **fragmentation**: 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: SA + CA

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
overview

- *also in the ICT WP:*
  "**4.1 SME initiative**", special SME-DCL call, 35 Meuro
- data is the crude oil of today's R&D and yet often too expensive for new or small actors
- ease development 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 **acquisition & aggregation**
  - digital **trading places**, open exchanges or commons
  - (experimental evidence of) **new or better services** resulting from combining, extending, repurposing… resources
- instruments: STR + CSA

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

ICT 2011-12
4.1 implementation

- **budget**: 35 Meuro for two domains (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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Towards FP8
timeplan

- EC proposal for financial perspectives after 2013: **mid-2011**
- FP8 proposals adopted by EC: **early 2012**
- agreement on the next financial framework: **mid 2013** (?)
- FP8 launch: **end 2013** (?)

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Towards FP8
some issues

- **better articulation of research & innovation**
  - FP8 ICT vs. CIP-II ICT after 2013
- **more strategic, focused research**
  - roadmap-based research
- **closer ties between research & infrastructure**
  - from physical to "knowledge" hubs
- **new implementation modes**
  - simple, light, fast, SME friendly …

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

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) [11/11 + 17/11]
- **mid-Jan 2011**:
  - close of Language Technologies call
- **Feb 2011:**
  - launch of SME call
- **close of SME call:**
  - Apr 2011 (1st stage, short proposals)
  - Sep 2011 (2nd stage)

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

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**
  - 80-100 submissions?
- **tell academics to bring vendors & users!**

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Thank you!

INFSO-E1@ec.europa.eu

Upcoming ICT-HLT calls & events
(under construction):

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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media