Language.Technologies
@wp.ict.eu

European Commission

INFSO - Information Society & Media

Digital Content & Cognitive Systems

Language Technologies & Machine Translation

infso-e1@ec.europa.eu

Brussels, 17 November 2010

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Agenda

- presentations
  - ICT Call 7, Q&A
  - ICT SME Call, Q&A
    - further questions if any
- short statements from the audience
- bilaterals with EC officials

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Foreword

- the aim is to teach computers how to understand & process **written & spoken human language(s)**
  - language is a powerful **medium**
  - for information, communication, interaction
- **(H)LT** – *(human) language technologies*
  - **several communities** & specialist groups
  - including but **not limited to linguistics**
    - statistics & machine learning
    - semantics & knowledge engineering
    - cognitive systems…

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

State of play
Opportunities

Research programme (FP7 ICT 2011-12)

- **LT** part of **Challenge 4** "*Technologies for Digital Content & Languages*"
- appears in **2 calls:**
  - **Call 7:** open Sept 2010,
    close Jan 2011 (1-step), 50 M
    - our "home", dedicated to LT
  - **SME-DCL:** open Feb 2011,
    close Sept 2011 (2-steps), 35 M
    - "open data", both content & language

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

How to proceed

- read the **work-programme**
- go through this presentation
- re-read the workprogramme
- ask for guidance & feedback
- (re)assess your project idea
- bear in mind that **success** will depend on
  - fitness v-a-v work-programme
  - potential in terms of impact
  - perceived quality

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
under Call 7

4.2 Language Technologies

- budget: 50 M
- instruments: IP, STR, CA+SA
- inquiries & pre-proposals until Fri 17 Dec
- closing date: 18 Jan 2011
- project start: Nov 2011 – Jan 2012

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
overview

- **3 research lines ("outcomes")**
  - a. (multilingual) content processing
  - b. information access & mining
  - c. natural spoken interaction
- each line, indeed every project is **multilingual**
- each line provides ample opportunities for
  - **ambitious** efforts
  - **cross-disciplinary** research
  - active co-operation with **users & vendors**

no cross-over between a., b. and c., stay within the line you've chosen

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
overview

- **basic common features**
  - **written and/or spoken language**, as required
  - **multilingual** (i.e. multiple in/out languages), where relevant cross-lingual ("translation")
  - handle conventional & **everyday language**
  - cope with **massive volumes** & diverse sources
  - cater for contextualisation & **personalisation**
  - technologies are **adaptive** (language, domain, task)
    - but **embedding & testing** within specific (demanding) application environments

be ambitious, be empirical, deliver useable results

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
instruments

- **no predefined budget allocation**
- **balanced mix** of projects
  - 50% STR (21 M)
  - 30% IP (13 M)
  - 20% open IP ⇔ STR (8 M)
- in addition, **coordination & support** actions
  - agenda: research roadmaps & partnerships
  - reuse: language resources & standards
  - exploitation:
    - technology transfer & market uptake
    - evaluation

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
timetable

- **selection**
  - April
- **negotiation**
  - from Easter (!) until Sept/Oct
- **project start**
  - ASAP after grant is awarded, in any case no later than Jan 2012
- **how many successful submissions?**
  - ~14 in total? incl. 2-3 IP's & 8-9 STR's

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
a closer look

- **3 project lines ("outcomes")**

  a. (multilingual) content processing

  b. information access & mining

  c. natural spoken interaction

- please come with
  - **fresh ideas**
  - **new participants**

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
project lines

a. multilingual content processing

- **human-to-human**; addresses the production (*outbound*) chain in a multilingual setting – authoring, translating & (web) publishing
  - language-encoded knowledge embedded in documents, databases, social media, web & audio-visual objects
- two project lines:

  (1) advance machine translation on several fronts
  - quality/fitness, self-learning, adaptation…
  - everyday language, x-lingual resources…

  (2) test & improve suitability (usability, effectiveness…) **of novel LT** in real-life conditions
- instruments: IP + STR

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.2
project lines

(1) is cutting edge

(2) is more applied & user driven:

- … within typical **production processes** and translation / localisation **workflows**, in **real-life multilingual settings**
- … **optimise & integrate** technologies within demanding application environments, **assess** their suitability & increase their potential
- … **field trials**… together with **user-centred & economic analyses**

high-quality domain MT; MT + social media;
MT + user feedback; MT + post-editing; MT + TM;
CAT; speech-2-speech translation…

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.2
project lines

b. information access & mining

- **human-to-information;** finding, categorizing, interpreting, correlating… digital content - the *inbound* chain
  - 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** (analytics)
  - **text mining**, diverse/multilingual sources
- instruments: STR

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.2
project lines

c. natural spoken interaction

- **human-to-computer**; progress towards richer, more spontaneous & robust man-machine interaction
  - it's **not** about robotics, nor technology-mediated inter-personal communication
- **"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 **proof-of-concept systems**, preferably within larger systems (e.g. mobile applications)
- instruments: IP + STR

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.2
cross-cutting actions

d. coordination & support – building on, extending & liaising with existing initiatives (positive overlaps but no duplication!)

- 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 **(re)usability** & **interoperability** of language data & tools by means of pooling & sharing
  - '**soft**' – **open standards** incl. methods, guides, best practices…
  - '**hard**' – **open repositories** of research, development & training resources…
- instruments: CA (small) + SA (bigger)

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Target languages

- **how many languages?**
  - depends on the proposal, its scale & depth of analysis; general rule: **3+**
- **what languages?**
  - EU official & working languages
    - including the national languages of non-EU countries participating in FP7 (e.g. Israel, Norway, Switzerland, Turkey…)
  - other languages of the EU member states
  - languages of EU trade partners

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

FAQs

- **how big?** STR 3+ M, IP 6+ M
- **how long?** up to 3 years in most cases
- **how many languages?** depends, 3+
- **how many partners?** as dictated by the project, as few as possible!
- **industry led?** a possibility, dep. on scale, hw/sw prerequisites, impact & timescale
- **user & commercial partners?** yes, whenever possible – you need both problems & market channels
- **use case(s)?** yes, always!
- … ? …

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Questions?

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1
SME-DCL call

4.1 SME initiative

- budget: 35 M
- instruments: STR (26 M), CA+SA (9 M)
- inquiries & pre-proposals from publication date until 31 Mar 2011
- 2 stages, submission deadlines:
  - 28 Apr 2011 (short proposal)
  - 28 Sept 2011 (full proposal, if passed 1st evaluation)
- go/nogo decision: early June
- selection: Nov 2011
- start: mid-2012

it's an experiment, if it works there will be more!

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.1
SME definition

what's an SME?

- 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 (notably independence)
- the **official definition** of SMEs can be found at

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

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.1
2-stage process

what's a "short" proposal?

- **part A** (forms with partners & resources) as in any normal ICT submission – for EC to check eligibility
- **part B** (narrative, 5 pages) is anonymous; it contains an outline description of the planned project:
  - **rationale**
  - **innovation**
  - **output**
  - **impact**
- no implementation details at this stage
- it is the potential & relevance of the "idea" that is going to be evaluated
- remember: at this stage you are **not** selected, you are simply invited to develop a full proposal

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1 rationale?

rationale is "Open Data"

- data is the crude oil of today's research & business, and yet often too expensive for new or small actors
- the idea is to "release the power of data", in practice
- … ease development & first-use deployment of novel data-intensive technologies by **high-tech SMEs**
- … so as to operate *large-scale* as corporations do
- … by **pooling** data sets & related data-processing tools
  - *knowledge (linked) data*, (a) + (b), objective 4.4
  - *language data*, (c) + (d), objective 4.2 (us!)
- instruments: STR & CA+SA

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Objective 4.1
tasks? STR

c. sharing language resources

- projects should address at least 2 of the following issues, #2 is mandatory
  1. acquire: make more effective the **acquisition/cleanup** of large-scale language resources with automated and/ or collaborative means
  2. share: contribute to **open exchanges** based upon the concerted pooling of resources
  3. reuse: show the **concrete impact** of using, combining or repurposing the above resources in a given **use context**
- we need **experimental evidence** of new or better technologies/services resulting from this process

Objective 4.1
sharing?

- **pooling & reuse** can be achieved in different ways
  - by purely **legal** means
    e.g. Creative Commons licences
  - by **legal & physical** means
    CC + storage/curation: open Web or existing multi-party repositories or other set-ups that will result from the concurrent CSA actions
  - **time-wise:**
    right from the outset, by the end of the project, within 12 months to preserve competitive advantage…
- suitable terms & conditions will be negotiated with the successful consortia

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1
tasks? CA+SA

d. building consensus & common services

- provide the "**glue**" between (i) existing & future projects, (ii) other players within the LT business & applied research communities

  1. soft element ("building consensus…")

  – mechanisms to mobilise the stakeholders, experiences & solutions in other domains, consensus on short & medium-term requirements, suitable schemes & platforms…

  2. hard element ("… and common services")

  – support services & pooling/trading facilities as defined by the partner projects & other stakeholders through the above mechanisms

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1
rightsizing?

- **focused STR projects**
  - up to 24 months
  - up to 2 M funding
- **compact STR consortia**
  - up to ~6 private/public partners
  - at least 2 SMEs (= not just SMEs!)
  - accounting for >30% of the total EU funding
- no a-priori constraints for **CA+SA's** other than common sense & available budget (4 M)

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1
which SMEs?

- **commercial LT developers/vendors**
  - text as well as speech
  - including but not limited to translation & localisation
  - incl. university spinoffs, start-ups…
- **providers of language services (LSP)**
  - with own LT capabilities
- **channels & integrators**
  - enterprise search & content management
  - text & content analytics
  - interactive media & edu-entertainement…

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Objective 4.1
timetable

- **selection**
  - November 2011
- **negotiation**
  - early 2012
- **project start**
  - ASAP after grant is awarded, in any case no later than July 2012
- **how many successful submissions?**
  - ~9 in total?

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

FAQs

- what accounts as a **language resource**? you tell us
- for what sort of **technology**? yours!
- how many **languages**? you decide, 3+
- how many **partners**? as dictated by the project, as few as possible; 4-5 in most cases?
- **industry** led? core tasks yes, not necessarily coordinator
- involvement of **commercial** partners: of course!
- can I revisit the **composition of the consortium** after the first evaluation? yes; evaluations are independent of each other
  - and yet still 2+ SMEs, >30% of the funding!

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Questions?

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

How about Language Resources?

- **compilation of x-lingual** LRs from the web & large-scale digital collections
  - under call 7 (a), within a broad-based MT project
- standards & platforms for **sharing** LRs
  - under call 7 (d)
- **SME-driven pooling & reuse** of LRs
  - under SME call, (c) & (d)
- creation, annotation… of **domain/task specific** LRs
  - call 7: within a relevant technology-driven project
  - SME call: under (c)

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Pre-proposals
Call 7

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

- **3 pages maximum**
  - rationale & problem area
  - contribution to WP esp. outcomes & impacts
  - consortium (outline)
  - scale – effort, duration, instrument
- to our functional mailbox
- before 17 December 2010

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Conclusion

wanted:

- **experts** for evaluations & project reviews
- **fresh ideas** & partnerships

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Thank you!

infso-e1@ec.europa.eu

ICT-LT events & projects:

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

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media