vicomtech

visual interaction & communication technologies

ICT2015 Event

Shape the multilingual Digital Single Market in e-commerce and join us at our networking Session

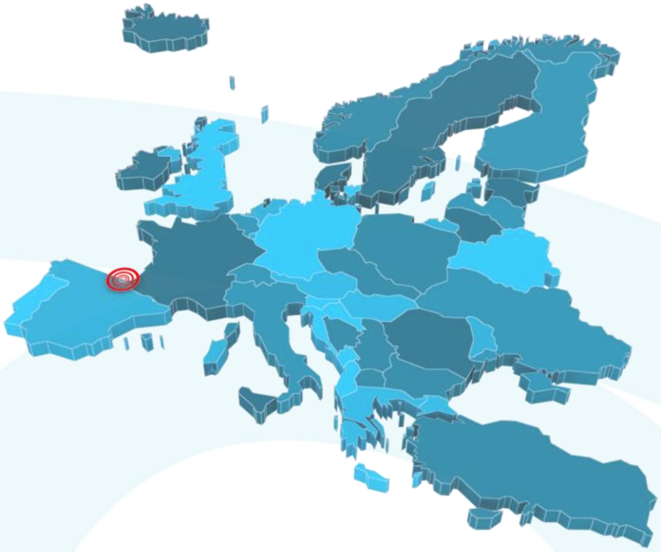
Language Technology in Big Data Applications: Challenges

Seán Gaines, Director of International Projects, 20151022

About us

Applied Research Centre, founded in 2001, specialising in

Computer Graphics, Visual Computing and Multimedia technologies



+105 Staff

(28% Ph.D.)

PhD., engineers, computer scientists, ….)

International Team

France, Germany, Colombia, Ireland, Slovakia …

Ph.D…. From (among others)

Imperial College London , Cambridge, Manchester, ETH Zurich, Bordeaux , Technical University of Darmstadt , Ohio State, University of Navarre, Basque Country, Deusto, UPM

Internal Organisation

Industry and Advanced Manufacturing

Intelligent transport systems and engineering

eHealth and Biomedical Applications

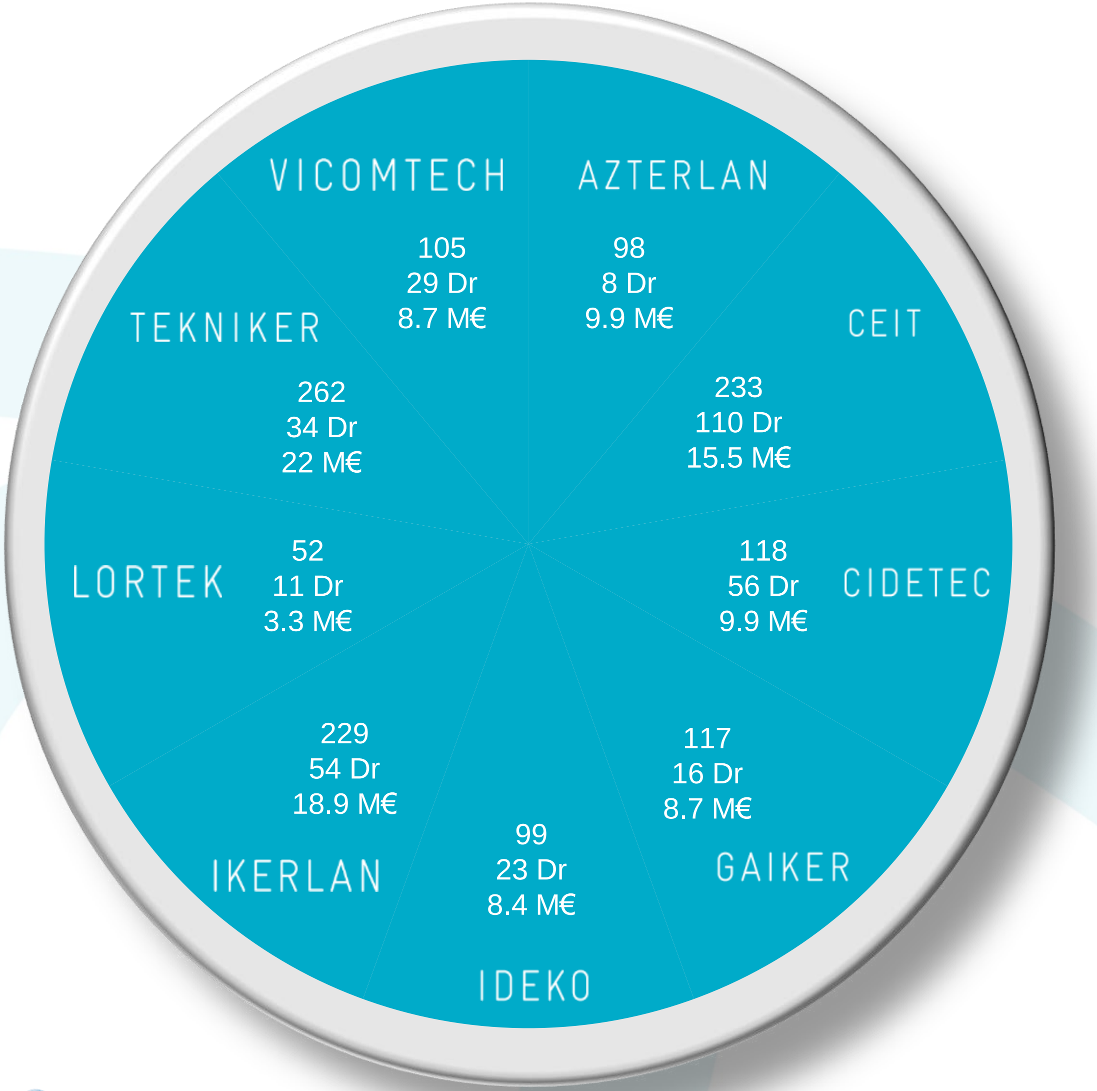
Digital Television and Multimedia Services

Interactive Computer Graphics

eTourism and Cultural Heritage

Speech and natural language technologies

IK4 Research Alliance



VICOMTECH	105	29 Dr	8.7 M€
AZTERLAN	98	8 Dr	9.9 M€
CEIT	233	110 Dr	15.5 M€
CIDETEC	118	56 Dr	9.9 M€
GAIKER	117	16 Dr	8.7 M€
IDEKO	99	23 Dr	8.4 M€
IKERLAN	229	54 Dr	18.9 M€
LORTEK	52	11 Dr	3.3 M€
TEKNIKER	262	34 Dr	22 M€

Key Data

IK4 Research Alliance

1,317

341 Dr

(26%)

105.3 M€

SAVAS

Acronym: SAVAS

Title: Sharing AudioVisual language resource for Automatic Subtitling

Role: Promoter and Coordinator

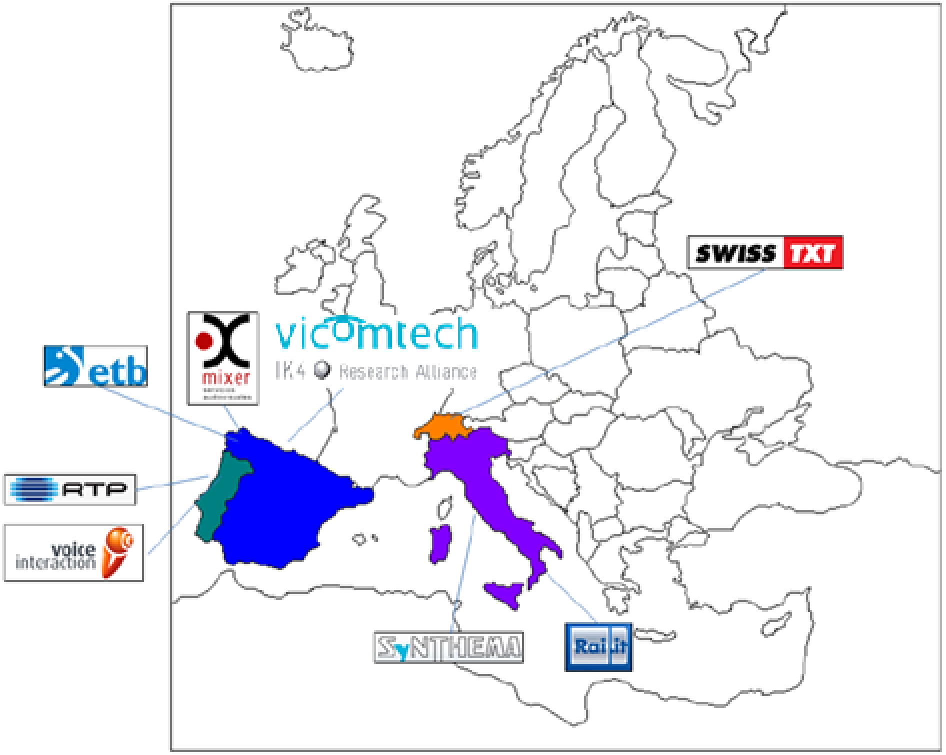
Grant Agreement: 296371

Type: STREP

Duration: 24 months

Call: FP7-ICT-2011-SME-DCL

Objective: ICT Call - SME initiative on Digital Content and Languages



SAVAS Sharing Audio Visual language resources for Automatic Subtitling

OpeNER

Acronym: OpeNER

Title: Open Polarity Enhanced Named Entity Recognition

Role: Promoter and Coordinator

Grant Agreement: 296451

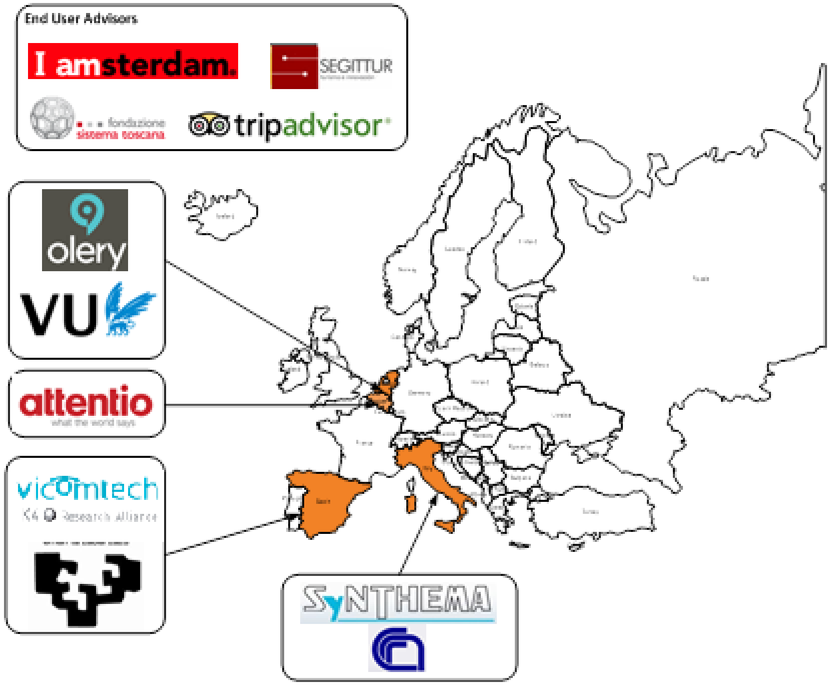
Type: STREP

Duration: 24 months

Call: FP7-ICT-2011-SME-DCL

Objective: ICT- 2011.4.1 SME Initiative on Digital Content and Languages, topic c)

Website: http://www.opener-project.org



OpeNER

SUMAT

Acronym: SUMAT

Title: An Open Service for SUbtitling by MAchine Translation

Role: Promoter and Coordinator

Grant Agreement: 270919

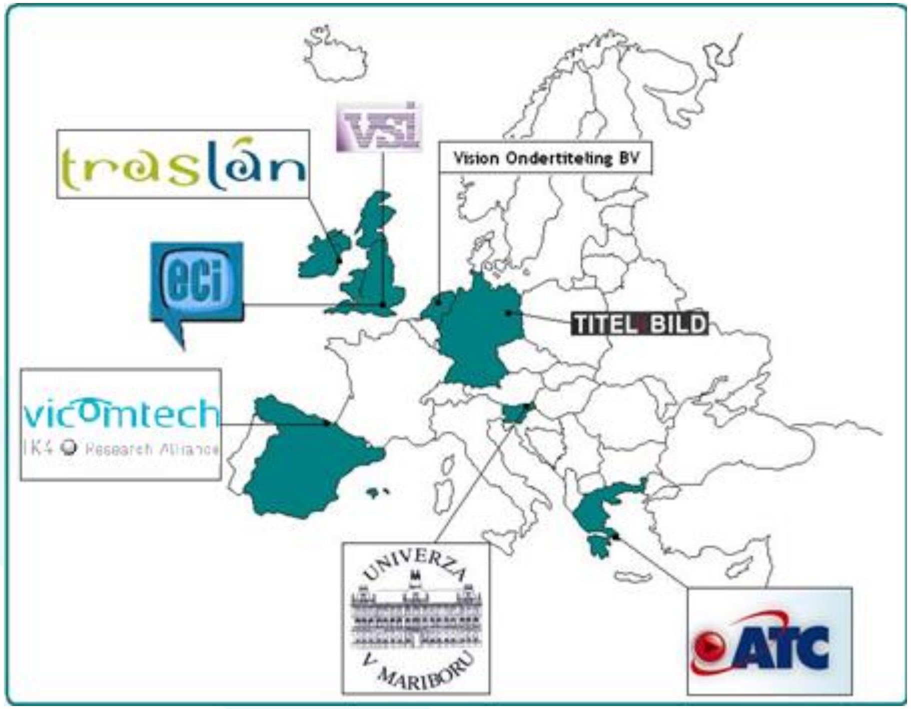
Type: Pilot Type B

Duration: 36 months

Call: CIP-ICT PSP-2010-4

Objective: 6.2. Multilingual online services

Web: http://www.sumat-project.eu/



CAPER

Acronym: CAPER

Title: Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime

Role: Promoter

Grant Agreement: 261712

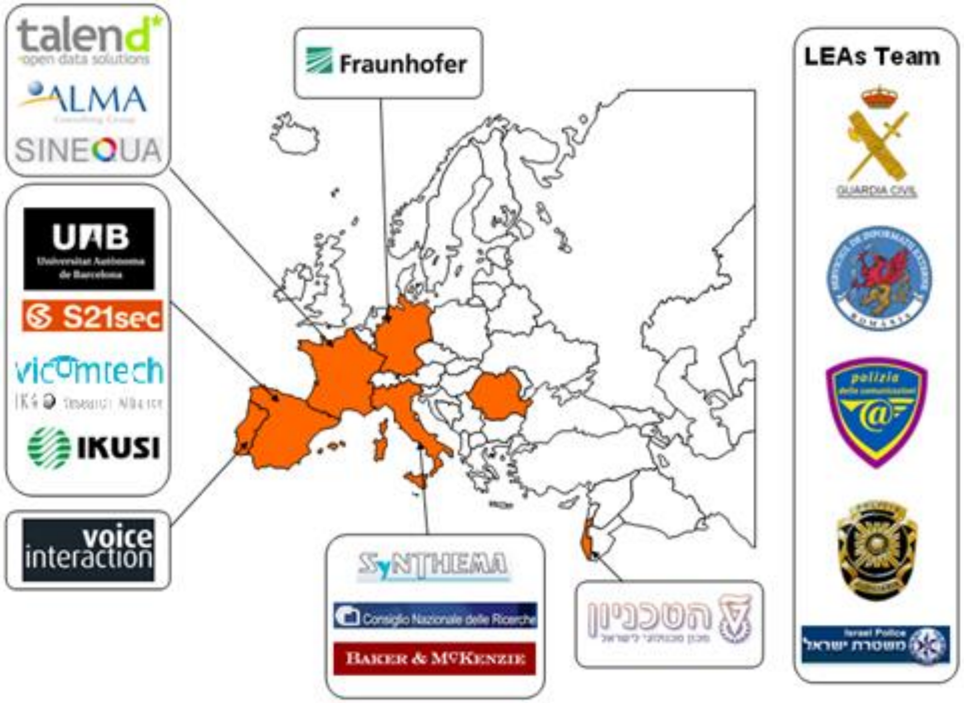
Type: CP – Large scale integrating project

Duration: 36 months

Call: FP7-SEC-2010-1

Objective: SEC-2010.1.2-1

Web: http://www.fp7-caper.eu/


talend open data solutions
ALMA
SINEQUA
UAB Universitat Autònoma de Barcelona
S21sec
vicomtech IK4
IKUSI
voice interaction
Fraunhofer
SYNTHEMA
Consiglio Nazionale delle Ricerche
BAKER & MCKENZIE
הטכניון מכון טכנולוגי לישראל
LEAs Team
GUARDIA CIVIL
polizia delle comunicazioni
Israel Police משטרת ישראל

LT in a Big Data World

- Big Data Application or Enabling Technology?
- BDVA SRIA just about includes LT
- EDF just about includes LT
- LT is not strongly shown as a growth area in Big Data
- Where does LT fit?

LT Market Challenges

- Strengthening Technology Transfer
- Maintaining digital representation of minority languages
- As always, Market I18n
- By 2020, tapping the:
  - 15 BN EUR market in HMI
  - 20 BN EUR market in Text Analytics
  - 30 BN EUR market in M/AT and NLP
- Current LT market of 30 BN EUR

LT Market Challenges

- SMEs have more need than resources for LT
- Fragmentation
- Lots of demand and money to satisfy it
- Shoestring budgets to do LT work
- High variation and lack of comparability in the market
- Infrastructure investment acts as a substantial barrier
- Sharing Capabilities
- Lack of consistent coverage per language

Language Adaptation Challenges

- Quality Data for Tools and Resources
- Uniformity in tools and implementation
- Benchmarking of tool performance
- Cross-Lingual linking
- Licenses, licenses, licences
- Cost barriers

Domain Adaptation Challenges

- Quality Data
- Perfect vs. Functional
- Many Domains, few have been addressed
- Benchmark, evaluation and testing
- Sectorial domains and technological domains – What are the key combinations?
- More open source repositories for data and tools
- Cost barriers

Who has the need?

- SMEs
- Freelancers
- Start-ups
- Public Administration
- Large Industry
- Etc. etc.
- But we still look at the same sectors over and over, with the same approach, over and over
- Who can drive the demand?

Barriers

- Access to Data and Tools
- Lack of uniformity in language tools
- Lack of uniformity in data pools
- Adoption of data models and format standards
- Ease of adoption by certain segments
- Placing value on data (and monetising it)
- Pieces are still missing to solve domain specific LT puzzles
- Multidisciplinary approaches

How to exploit?

- *X*aaS
  - – Platform, Software, etc. as a service
- Shrink Wrapped
- SDK
- Bespoke Services and Products
- Innovative Business Models
- Pools and Infrastructure
- “Dumping” it to Open Source

Transparent Open Source, Licensing and Commercialisation schemes

- Variety and danger in License types
- Exploitation Models
  - Transaction Based Charges
  - Mixed Models
  - Direct Sale of Licenses
  - User Types Model
  - “Give and Take” models
  - Subscription Model
  - Pay on Deploy Model
  - Hybrid Models
  - Open Source with Ring-fencing

Technical Challenges

- Current LT needs a large technological step forward need before demand can be satisfied
- In specific bounded domains the technology is sufficient to address industrial needs, but others are not, leading to niching
- Big Data is a double edged sword:
  - 3Vs of data complicate the LT challenge
  - Big Data needs LT to address its own challenges

Technical Challenges

- Limited resources should be spent on:
  - Research: Filling the gaps in undone work, making it easily and cheaply available, addressing under or incompletely resourced, and minority languages
  - Industry: Spend on the LT enabled Services or Products, not on building the technology
- Knowledge of standards, not the standards themselves
- Create a "pull" demand in the LT market
- Organisation of LT providers and research
- Data silos are not the only silos

www.vicomtech.org
info@vicomtech.org