Antonio Zampolli (1937-2003)

Antonio Zampolli (1937-2003)

Brief Overview of recent activities in Europe

Khalid CHOUKRI
ELRA/ELDA
55 Rue Brillat-Savarin, F-75013 Paris, France
Tel. +33 1 43 13 33 33 -- Fax. +33 1 43 13 33 30
Email: choukri@elda.fr
Web: http://www.elda.fr/

Outline

Major Speech resources collections

Recent National activities .. The case of France

The European HLT landscape – HLT benchmark

(results of EuroMap and the achievements of Enabler)

The new European R&D Framework program

Some shortlisted projects

A New instrument for Coordination LangNet

LangTech2003 (Paris), LREC2004 (Lisbon)

Centralized Not-for-profit organization

for the collection, distribution, and validation of Language Resources and tools.

Operational agency: ELDA:

Evaluation & Language Resources Distribution Agency

An Association of users of Language Resou

A Repository Center:

- Technical & Logistic issues
- Commercial issues (prices, fees, royalties)
- Legal issues (Licensing, IPR)
- Information Dissemination

Infrastructure for the evaluation of Human Language Technol

providing resources, tools, methodologies, logistics,

Exit strategies / Capitalization on evaluation packages

European Funding(s)

Public Funding

- Commission of the European Union(e.g. R&D
- National agencies & Authorities

Private Funding

Brief Overview of recent activities in Europe European Union Level

Some Projects within FP5 and previous FPs …. Related to Cocosda concern

Resources production: Speechdat Family

& Some other projects (TC-STAR-P, NESPOLE!, FAME, etc.)

Specifications of new types of resources: Natural Interaction and MultiMo

within ISLE (International Standards for Language Engineering) project (ov

Production of MM resources .. M4: MultiModal Meeting Manager

Standards/metadata: A major project started early 2003 INTERA

Coordination: ENABLER ,

Partnership between Europe and Mediterranean countries NEMLAR

Information gathering & Dissemination /benchmarking : Euromap

SpeechDat Family

SpeechDat(M) --- Fixed Telephone network -- 1K Speakers

SpeechDat-II Fixed, Mobile, 1-5Kspeakers

SpeechDat-II Speaker Verification

SpeechDat-E (CEE - Polish Czech Slovak Russian Hungarian)

SALA (Speech Across Latin America) **and** Now SALA-II

SpeechDat-Car (inc. cellular)

SpeeCon (Consumer products)

Orientel

SpeechDat

SpeechDat-Car

EAST SpeechDat

SALA SpeechDat across Latin America

SpeechDat Family: SALA-I STATUS OF DATABASE VALIDATIONS

Language	Owner	Validate	availability
Brasil	Philips & CSELT	16 Aug. 2000!	
Mexico	Vocalis	??????	
Venezuela	UPC	10 November 2000!	Yes
Colombia	Siemens	8 June 2000!	Yes
Argentina	Lucent	17 August 2001!	
Chile	Temic	14 December 2001!	

SALA II cellular/Mobile Network (1000 speakers)

Partner	Latin America	US and Canada
ATLAS	Venezuela	
Loquendo	Chile,	US English South
Lucent,	Argentina	
Microsoft	Peru	US English North
NSC	Mexico	US English Midland
Philips	Brazil	US Spanish West
Siemens	Colombia	US English West
Telisma	Costa Rica	Canadian French

SALA II cellular/Mobile Network (1000 speakers)

Partner	Latin America	US and Canada
ATLAS	Venezuela	US Spanish East
Loquendo	Chile	English USA
Microsoft	Peru	English USA
NSC	Mexico	English USA
Philips	Brazil	US Spanish West
Siemens	Argentina	English USA
Telisma	Costa Rica	Canadian French

SpeeCon Project

Participant number	Participant Name	Participant short name	Country
1	Siemens Aktiengesellschaft	Siemens	Germany
2	Ericsson Eurolab Deutschland GmbH	EEDN	Germany
3	IBM Deutschland Entwicklung GmbH	IBM	Germany
4	Lernout & Hauspie Speech Products NV	L&H	Belgium
5	Matra Nortel Communications	Matra	France
6	Nokia Corporation	Nokia	Finland
7	Philips Speech Processing Aachen Zweigniederlassung der Philips GmbH	Philips	Germany
8	Sony International (Europe) GmbH	Sony	Germany
9	TEMIC TELEFUNKEN microelectronic GmbH	TEMIC	Germany
10	DaimlerChrysler AG	DCAG	Germany

Dialectal zone	*Language*	*Region*	*Remarks*
Esl_ES	Spanish	Spain	(excluding Latin America)
Rus_RU [1)]	Russian	Russia	
Ita_IT	Italian	Italy	
Sve_SE_FI	Swedish	Sweden and Finland	
Deu_DE_AT	German	Germany and Austria	(excluding e.g. Belgium, Luxembourg, Switzerland)
Eng_GB	English	United Kingdom	
Dan_DK	Danish	Denmark	
Dut_BE	Dutch	Belgium	
Fra_CA	French	Canada	
Fra_FR	French	France	(excluding e.g. Belgium, Luxembourg, Switzerland)
Fin_FI	Finnish	Finland	
Zho_CN_HK	Mandarin	P. R. China (incl. Hongkong)	(excluding e.g. Taiwan)
Dut_NL	Dutch	The Netherlands	
Jpn_JP	Japanese	Japan	
Pol_PL	Polish	Poland	
Por_PT	Portuguese	Portugal	(excluding Brazil)
Deu_CH	German	Switzerland	
Eng_US	English	USA	(excluding e.g. Canada)

Partner	Language	Region	Remarks
Siemens	Spanish	Spain	(excluding Latin America)
	Russian	Russia	-
Ericsson	Italian	Italy	-
	Swedish	Sweden and Finland	-
Herterkom	Hungarian	Hungary	-
	Czech	Czechia	-
IBM	German	Germany and Austria	(excluding e.g. Belgium, Luxembourg, Switzerland)
	English	United Kingdom	-
NSC	Hebrew	Israel	-
	French	France	(excluding e.g. Belgium, Luxembourg, Switzerland)
Nokia	Finnish	Finland	-
	Mandarin	P. R. China (incl. Hongkong)	(excluding e.g. Taiwan)
Philips	Dutch	The Netherlands	-
	Japanese	Japan	-
Scansoft	Danish	Denmark	-
	Dutch	Belgium	-
Sony	Polish	Poland	-
	Portuguese	Portugal	(excluding Brazil)
TEMIC	German	Switzerland	-
	English	USA	(excluding e.g. Canada)
External non-funded partner:			
Microsoft	Cantonese	China and Hongkong	-
	Thai	Thailand	-
Panasonic	Spanish	USA	-
	Mandarin	Taiwan	-

SpeechDat Family: OrienTel

Multilingual access to interactive communication services for the Mediterranean and the Middle East

7 linguistic regions 10 “OrienTe”l countries 23 databases

SpeechDat Family: OrienTel

Linguistic affiliation	*OrienTel* countries	Languages covered
Mahgreb Arabic (excluding Algeria and parts of Libya)	Morocco	Standard Arabic Colloquial Moroccan Arabic French
	Tunisia	Standard Arabic Colloquial Tunisian Arabic French
Egyptian Arabic (excluding parts of Libya)	Egypt	Standard Arabic Colloquial Egyptian Arabic English
Levantine Arabic (excluding Syria, Lebanon and Jordan)	Israel and Palestine Authorities	Hebrew Standard Arabic Coll. South Levantine Arabic
Gulf Arabic (excluding Kuwait, Bahrain, Qatar, Oman and Yemen)	United Arab Emirates	Standard Arabic Colloquial Gulf Arabic English
	Saudi Arabia	Standard Arabic Colloquial Gulf Arabic English
Cypriote Greek	Cyprus	Greek English
Hebrew	Israel	Hebrew
Turkish	Turkey, Germany for German	Turkish German

SpeechDat Family: SALA - II

Phase II Cellular/Mobile Network	
Latin America	**US and Canada**
Mexico	US English North East
Argentina	
Chile*	US Spanish East
Brazil	English South West or US Spanish West.
Colombia	US English North West
Venezuela	
Costa Rica*	US English South East
Peru*	Canadian American English

US English North West
US English South West
US English North East
US English South East
US Spanish East (Caribbean var
US Spanish West (Mexican vari
Canadian British English
Canadian American English
Canadian French

M4 :MultiModal Meeting Manager (03/2003 – 3Y)

- The overall objective of the project is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings. The archived meetings will have taken place in a room equipped with multimodal sensors.
- For each meeting, audio, video, textual, and (possibly) interaction information will be available. Audio information will come from close talking and distant microphones, as well as binaural recordings. Video information will come from multiple cameras. While the video and audio information will form several streams of data generated during the meeting, the textual information---the agenda, discussion papers, text of slides---will be pre-generated and can be used to guide the automatic structuring of the meeting. The interaction stream consists of any information that can help in analysing events within the meeting, for example, mouse tracking from a PC-based presentation or laser pointing information.

M4 :MultiModal Meeting Manager (Objectives)

- Development of a ``smart'' meeting room, collection and annotation of a multimodal meetings database.
- Analysis and processing of the audio and video streams:
  - Robust conversational speech recognition, to produce a word-level transcription;
  - Recognition of gestures and actions;
  - Multimodal identification of intent and emotion;
  - Multimodal person identification;
  - Source localization and tracking.
- Although the technologies addressed here are imperfectly developed, they are established firmly enough to warrant their use in combination. Integration of multiple sensory inputs is a challenging problem at the early stages of investigation.
- Integration and structuring using the output of the various recognizers and analyses:
  - Specification of a flexible intelligent information management framework;
  - Models for the integration of multimodal streams, including statistical models for asynchronous multiple streams, multimodal syntax and multisource decoding;
  - Summarization of a meeting, or a meeting segment; this could take various forms such as a textual precis or a set of video key frames;
  - Multimodal information extraction and cross-lingual retrieval/browsing across the archive.
- Construction of a demonstrator system for browsing and accessing information

M4 :MultiModal Meeting Manager

- **Public Deliverables**
- **D1.1 Specification of smart room environment and data collection and annotation protocols**
- **D1.2Collection and annotation of meeting room dataMarch 2004**
- **D2.2Final report on multimodal recognizersMarch 2005**
- **D3.3Final report on multimodal information access March 2005**
- **D4.3Report on final demonstrator March 2005**

Other resources' oriented projects

C-Oral-Rom : Conversational Speech

Roman Languages: French, Italian, Spanish, Portuguese

"Comparable" data

Net-DC: BroadCast News speech Corpus

Brief Overview of recent activities in Europe

European Union Level

A major project within MLIS …. Related to Cocosda concerns

NETWORK-DC: Network of international & regional Data Centers

Partners: ELRA, SPEX & LDC

Others (GSK,…) welcome to join

Some Projects within FP6 …. Related to Cocosda concerns

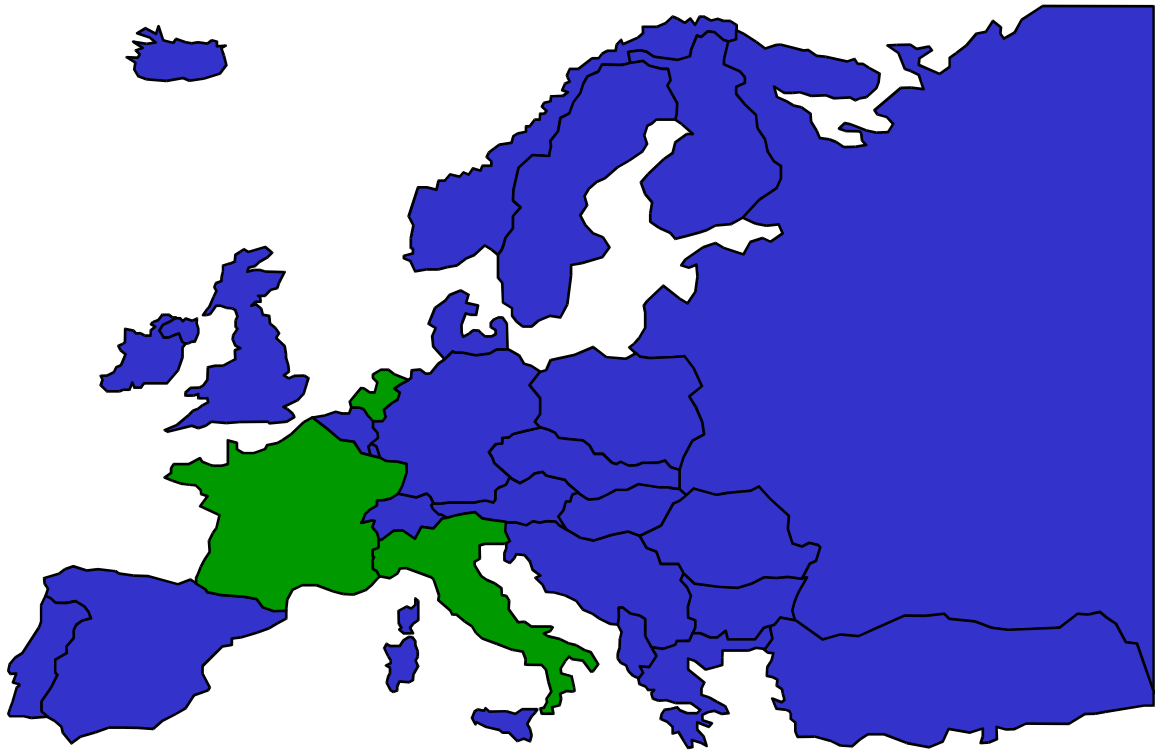
THREE IP (Integrated Projects)

AMI

CHIL

TC-STAR

Brief Overview of recent activities
National level



Examples of
National Projects/programs

OVER Nine National projects, among which :

Netherlands & Belgium: Continue Now Release 5

Data Available via ELRA, Release of April2002

France: Action Techno-Langue

Italy : Infrastruttura nazionale per le risorse

linguistiche nel settore del trattamento automatico della lingua naturale parlata e scritta

Norway : **Norwegian Language Bank**

Dutch & Flemish

Release 1 (March 2000)

- 62 hours speech samples orthographically transcribed (615,000 words), 90,000 words enriched with Part-of-Speech tags;
- annotation CD with first version of PRAAT (annotation tool) and first version of documentation (in Dutch) among which relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment) (information on the speakers in anonymous form);

Release 2 (October 2000)

- over 150 hours of speech samples, orthographically transcribed (over 1,500,000 words), approximately 750,000 words enriched with Part-of-Speech tags;
- annotation CD with annotation protocols and relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment) is available (information on the speaker in anonymous form);

Release 3 (April 2001)

- more orthographically data enriched with Part-of-Speech tags;
- the first broad phonetic transcriptions, word alignments, syntactic annotations, lexicon link-up will be available;
- annotation CD with documentation among which relevant information on the speakers (e.g. gender, age, socio-economic class) and samples (e.g. recording conditions, the equipment);

this release encompasses the first version of Corex, the exploitation tool.

Examples of National Projects/programs

Norway : Norwegian Language Bank

language technology resources in Norway

Launch conference 24-25 October 2002 (Bergen, Norway):

The language bank will contain three types of data spoken data, text and lexical resources

It will be organized as a foundation with state ownership,

The estimated budget is about NOK 100 million, (12 M€)

Examples of National Projects/programs

Italy : Infrastruttura nazionale per le risorse

linguistiche nel settore del trattamento automatico della lingua naturale parlata e scritta

•ItalWordNet (~50.000 entries).

•Corpus di italiano parlato --- 100 Hours of speech consisting of :

a) 10h Radio-TV broadcast data (notiziari, interviste, *talk show*),
b) 60h Map task like collection
c) 5h Lab data for lexical coverage
d) 10h telephone conversational speech
e) 10h Domain specific (finances, touristic information etc.)

•Annotated dialogues for speech interfaces (H-H and H-M interactions)
(Dialoghi annotati per applicazioni di interfacce vocali avanzate)
450 dialogues annotated at all levels (morphological … Prosody…Semantics ….)

Example of France

National Projects/programs

Technolangue Action

With Contribution from J. Mariani

TechnoLangue » Call

- Language resources
  - Basic Language Processing Tools (Open Source)
  - Spoken/written data (corpus, dictionaries, terminological data…)
  - Production, validation, distribution (incl. legal, economical aspects)
  - For a large use by a large community (education, training…)
- Evaluation
  - Technology (evaluation campaign)
  - Applications (evaluation toolkits)
  - Methodology (metrics / protocols)
- Norms & standards
  - Shared effort to improve French participation
- Technological survey
  - In relationship with on-going actions (Euromap...)

Part 1: Language Resources

- Stimulate the production and the distribution of language resources for :
  - answering minimal needs (*Basic LAnguage Resource Kit*) for the french language ;
  - promoting resources reusabilty ;
  - supporting research ;
  - helping industrial applications development ;
  - decreasing the cost of entering the sector for new comers
- Should include the French language, Possibly in connection with other languages

Part 1: Language Resources

- Spoken and written data :
  - oral corpus, pronunciation lexicons, etc.
  - databases for speech synthesis ;
  - monolingual and multilingual text corpus (parallel, comparable...) ;
  - lexicons, terminology, grammars,...
  - Lexical semantic resources : ontologies, thesauri,...
  - Multimodal corpus,...etc
- Basic sofware tools :
  - morphosyntactic taggers, syntactic parsers, semantic tools,
  - teminology extractors,
  - language identifiers,
  - corpus annotations tools,
  - lemmatizers,… etc.

Part 1: Language Resources

- Encourage and facilitate the use of those resources
  - Putting them in new (young) user hands
  - Same approach as for GUIs : “VUIs”
  - Language Technology Kits with “User’s guide”
    - Distribution towards specialized education entities (NLP, Document Engineering…) and more largely towards training centers (Universities, Technical Universities, Engineering schools...)
    - While insuring a feedback from experience
  - Open Source software economical model

Results: 26 selected projects:

- 8 on Language resources: incl.
  - Speech (Adulte (Neologos) + Children)
  - BLARK (Cf BNC), Fr-En, G, Sp, It, Arabic dictionaries
  - Specialized (aerospace, automotive…), proper names dictionaries
  - Aligned corpus (7 novels 19th century litterature in 4 languages)
- 6 on Tools (Open source)
  - Lemmatizer, Chunker, Guesser, Tagger, Parser, Speaker recogn., Topic & NE detector, summarizer, term. extractor, Search engine...
- 3 on Standards (Spoken / Written)
- 1 on Technological survey (Portal)
- 8 on Evaluation : 7 on technology, 1 on usage evaluation

Technology Evaluation

- Written language
  - Machine translation
  - Text alignment
  - Syntactic parsing
  - Information query
- Spoken Language
  - Speech transcription / indexing (incl. Named Entity)
  - Speech synthesis
  - Spoken dialog
  - *Possibly one on speaker Id….*

Results

- 52 proposals submitted
  - Total proposal costs : 35,9 M€
  - Total requested support : 21,7 M€
  - Clustering within each of the 4 topics
- 26 projects selected
- 173 participations, 94 participants :
  - 33 industry
  - 39 public research
  - 11 other (Associations, CEA, DGA…)
  - 11 foreign (Bell Labs, NII, EPFL, LATL…)
- Budget : 6,2 M€

« TechnoLangue » Call

- **International cooperation**
  - Cooperation mechanisms within TechnoLangue
    - foreign entities may participate in the projects
    - financing from their own funds
  - Future cooperation among similar national programs
    - EU Countries (Italy, Germany, Norway, Spain, Greece, The Netherlands, Switzerland…)
    - Prepare the construction of the European Research Area (ERA)
      - **The EC supports the coordination and generic technologies cost**
      - **Each country supports the cost for covering its language (s): specific technology development/adaptation: (annnotated) corpus (spoken/written), lexicon (incl. pronun.), dictionaries...**

Example of NORWAY

National Projects/programs

Norway : Norwegian Language Bank

language technology resources in Norway

Launch conference 24-25 October 2002 (Bergen, Norway):

The language bank will contain three types of data spoken data, text and lexical resources

It will be organized as a foundation with state ownership,

The estimated budget is about NOK 100 million, (12 M€)

AURORA (Speech distributed recognition)

Set up to establish a worldwide standard for the feature extraction software in a DSR (Distributed Speech Recognition) system:

(i) Evaluation of algorithms for front-end feature extraction algorithms in background noise

(ii) Evaluation and comparison of the performance of noise robust speech recognition algorithms.

-- A new Initiative

Identification of existing resources (Universal Catalogue)

The Basics (e.g. Standards, tools, evaluation procedures, …)

Survey of existing national activities

Fostering common research and compatibility of LR

Suggestion for and contribution to international

Extension foreseen/ Planne

Next meeting Berlin 25 September -- Cocosda representa

The ENABLER Mission

- Language Resources (LR) & Evaluation: central componen of the **"linguistic infrastructure"**
- LR supported by national funding in **National Projects** in Europe
- Principle of **subsidiarity** at the basis of many initiatives
- **Availability of LR** also a "sensitive" issue, touching the sphere of linguistic and cultural identity, but also with economical and political implications

ENABLER, setting up **a Network of National initiatives**, aims at "enabling" the realisation of a urgently needed **cooperative framework**

ENABLER Main Goals

- exchange **information** & forge **links** betw. national activities
- create a structured **repertory** of organisational and technical info
- promote the adoption of **de facto standards, best practices** & foster interoperability of results,

- discuss **innovative research needs**,
- formulate a **common agenda of medium- & long-term research priorities, … RoadMap(s)**
- identify **synergies on common research issues**

in particular in view of the construction of extended **multilingual LR**

ENABLER / ELSNET Workshop

International Roadmap for Language Resources

Paris, 28th-29th August 2003

ENABLER

European National Activities for Basic Language Resources

Designed and coordinated by Antonio Zampolli

Thematic Network

Action Line: IST-2000-3.5.1

NEMLAR (2003-2005)
www.nemlar.org

A Network for Euro-Mediterranean LAnguage Resource and human language technology development and support

to establish a network of partner centres of best practice in Arabic and other southern Mediterranean languagse processing dedicated to

- surveying the state of the art on language resource needs,
- establishing development priorities,
- validating the interoperability of components and standards,
- and developing a minimum set of language resources in order to enable linguistic diversity in the southern and eastern Mediterranean region

1st Survey at **www.nemlar.org**

Conference by Fall 2004

NEMLAR (2003-2005 www.nemlar.org)

1. **Center for Sprogteknologi (CST) (Denmark)**
2. **European Language Resources Distribution Agency (ELDA) (France)**
3. **Utrecht University, Faculty of Arts (NL)**
4. **University of Lyon (France)** Department of Arabic, Faculty of Languages,
5. **The Faculty of Information Technology, Amman University (Jordan)**
6. **ENSIAS (Morroco)**
7. **RDI (Egypt)**

Local Experts

8. **Commissariat à l' Energie Atomique(France)**
9. **Centre Nationale de la Recherce Scientifique Rhône-Alpes**
10. **Institute for Language and Speech Processing (ILSP) Greece**
11. **The Open University (UK)**
12. **University of Balamand (Lebanon)**
13. **Sotelet-IT (Tunisia)**

Would like to be associated?

HLT CENTRAL

http://www.hltcentral.org

EUROMAP - HOPE

is a knowledge building and dissemination project

whose main goal is to

raise awareness about

the market readiness and potential benefits of

Human Language Technologies (HLT)

among appropriate market players in the information society.

EuroMap Collaborative work

- EUROMAP (benchmarking)
  - All EUROMAP partners
  - Reports (Full Report + Summary)

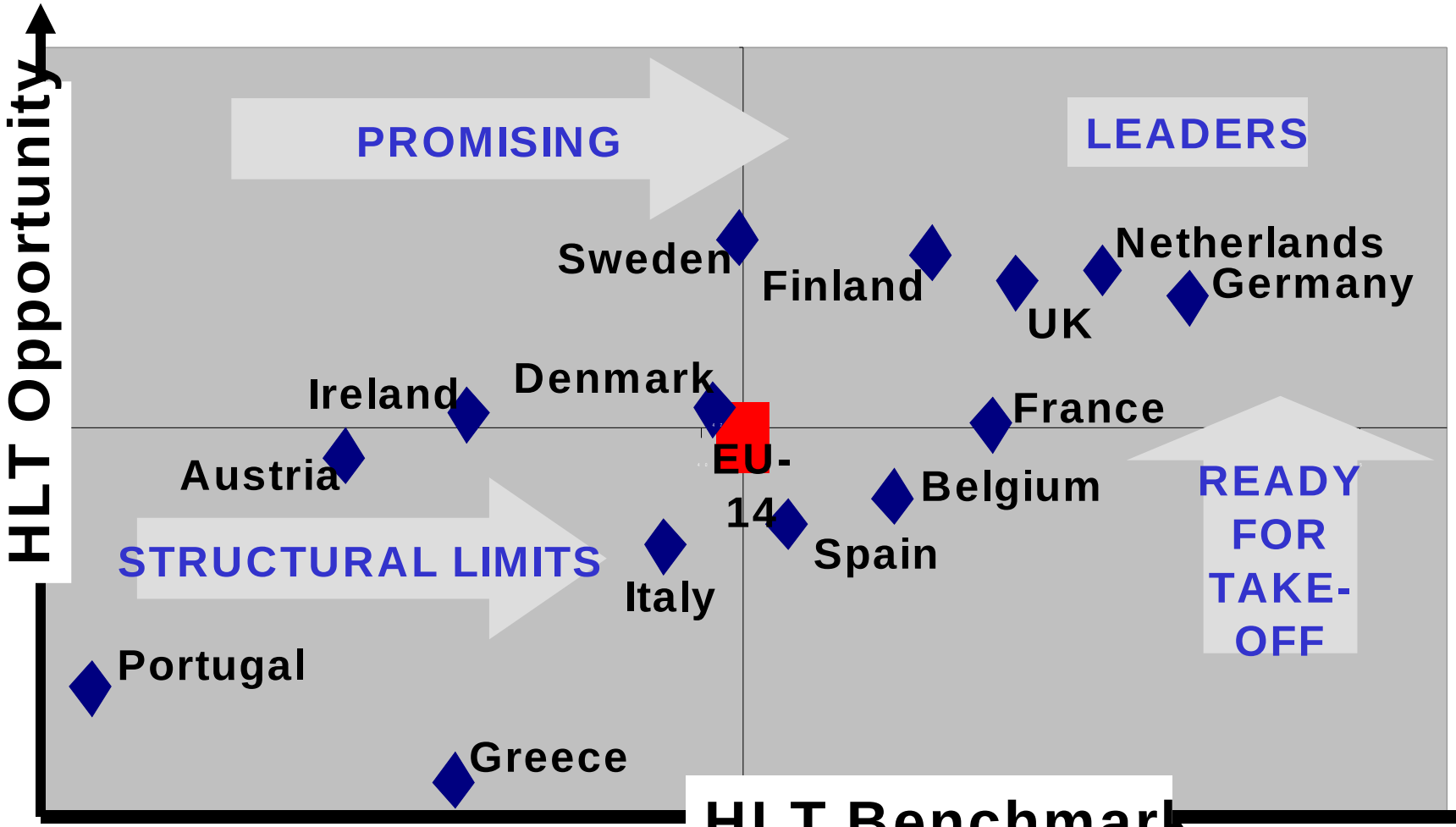
    On the web + Hardcopies

Andrew Joscelyne, Rose Lockwood

EUROMAP Benchmarking

- State-of-the art of HLT research and take-up in Europe (EU)
- Background for each country
  - Research centres
  - Suppliers
  - National research policies
  - Market analyses

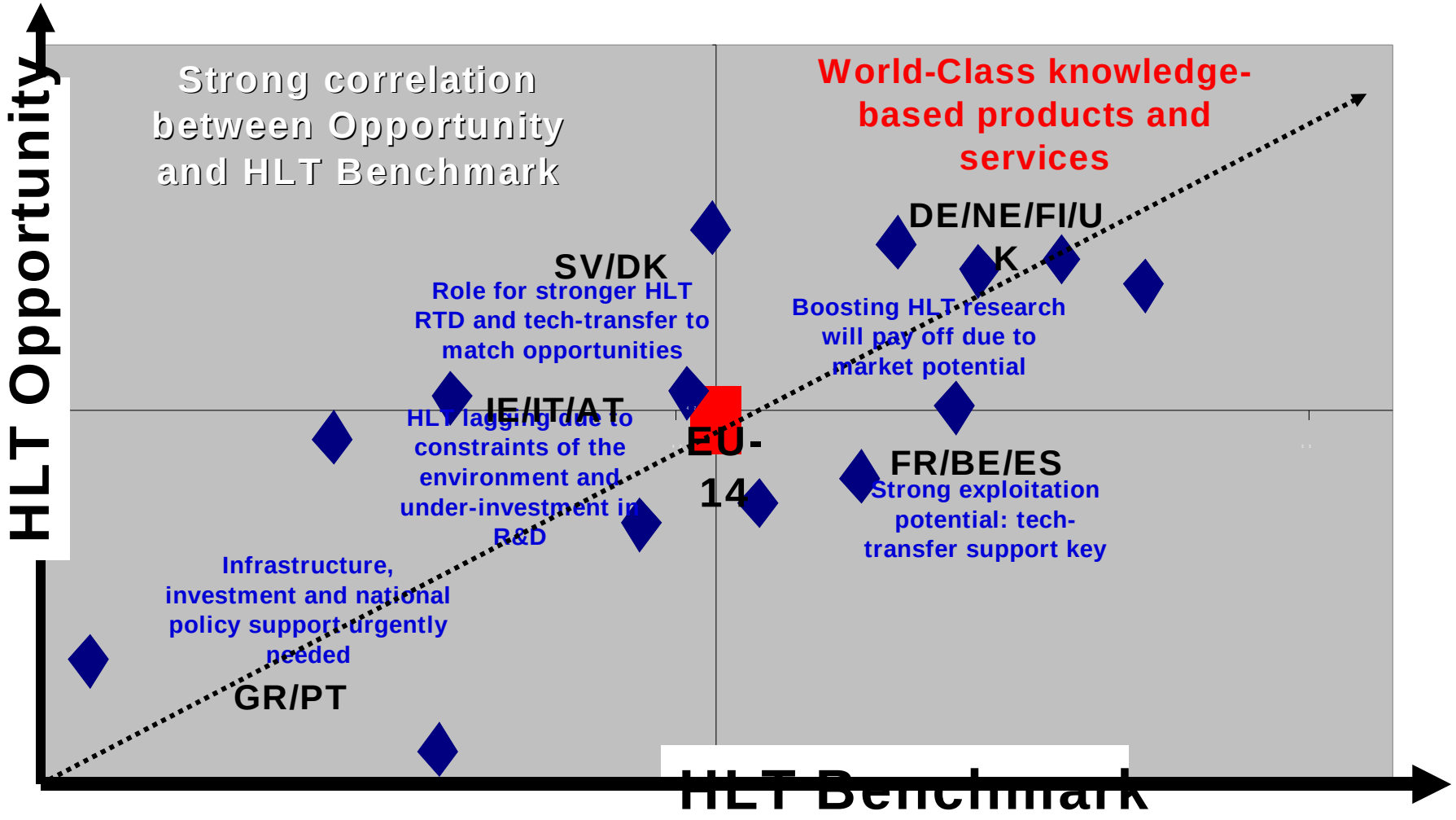
EUROMAP HLT Scorecard



Analysis for each country

- HLT research level and market take-up are correlated
- DE, NL, UK are the 'leaders'
- For each country the background has been compared with the analysis results
- DE, NL, UK have had a significant and steady public investment – long-term strategy

Guide to Action in each country



The need for LR

- If EU is to become truly multilingual
  - All languages should be covered at a minimal level
  - This requires LR and components for all languages

- Otherwise we will have a two-speed EU
  - The Leaders are constantly embedding new advanced systems for 'strategic' languages
  - The Rest attempts to ensure baseline coverage for the less strategic languages

Which industries?

- LRs are needed for
  - Language industry (producers)
  - Integrators (producers)
  - Translation industry (users)
  - Content industry in general (users)

Enabler Collaborative work

- ENABLER (needs survey)
  - Valerie Mapelli, Mahtab Nikkhou, Khalid Choukri, ELDA
  - Claus Povlsen, CST

ENABLER focus

- Focus on Producers and Integrators, not users
- Both large and small
- European (mostly)

Summary for WLR (4)

- Wish list
  - LRs for less used languages (less commercial interest in their development)
  - Domain-specific multilingual lexica/term bases
  - Domain specific corpora
  - Annotated corpora or treebanks
  - Semantic ontologies
  - Monolingual dict. for general language !

Summary for SLR (5)

- Wish list
  - Tools to manipulate LRs (parsers etc)
  - Broadcast news
  - GSM in close talk mode
  - SpeechDat model should be used for 'neglected' languages
  - Spontaneous task oriented speech (travel, movie)
  - Labelled data

Survey Conclusion

- ENABLER investigation confirms
  - Not sufficient language resources available
  - In particular not for less used languages
  - Language industries could do much more if more good quality LRs were available
- In line with the EUROMAP report
- Infrastructural funds are needed

Center for Sprogteknologi	CST	DK
VDI/VDE Technologiezentrum Informationstechnik GmbH	VDI/VDE-IT	DE
VIKOP Verein für Internationale Forschungs-Technologie und Bildungskooperation	BIT	AT
Instituto Cervantes	IC	ES
Scientific Computing Ltd.	CSC	FI
Consorzio Pisa Ricerche	CPR	IT
Arax Limited	Arax	UK
European Language Resources Distribution Agency	ELDA	FR
University of Brighton	ITRI	UK
Institute for Language and Speech	ILSP	GR
Nederlandse Taalunie	NTU	NL
Central Laboratory for Parallel Processing	CLPP	BG

LangNet – ERA-Net

ERA-NET

- 6th Framework Program (2002-2006)
- Construction of the European Research Area (ERA)
  - Article 169
  - ERA-NET (may prepare Article 169…)
- Cooperation & coordination of national or regional research and innovation activities (i.e. programs)
    - Exchange of information and best practices (projects, priorities, evaluation process, management…)
    - Strategic activities (complementarities,multinational scheme,barriers)
    - Implementation joint activities (clustering,joint evaluation,testbeds)
    - Transnational research activities (joint programs, calls, funding...)

ERA-NET

- 148 M€ total FP6 budget (2002-2006)
- EC funds the coordination costs, no R&D
- First deadline : June 3, 2003 (prep. May 17) - 24 M€
- Eligibility: Legal entities from at least 3 EM / AS
  - Public bodies financing/managing research at national/regional level
  - Other organizations doing the same (charities)
  - European bodies (EEIG…(may be sole as a group of public bodies))
- Coordinating Actions (CA)
  - 3 M€ max. over 5 years max. - 100% costs (incl. subcontracts)
- Specific Support Action (SSA)
  - Preparation of CA: 200 K€ max. over 1 year max.

LangNet

- Language as a specific issue for Europe
- Effort shared between the EC and EU member states
- Infrastructural nature (Cf TechnoLangue)
  - Language Resources (Cf Enabler ?)
  - Language Technology evaluation
  - Norms and standards
  - Survey (Cf Euromap ?)
- Extension in the future to a larger program including EC FP6 projects (TCstar…)

LangNet

- Identify EU Countries having similar programs
  - France (TechnoLangue)
  - Italy, Germany, Norway, Finland, Denmark, TheNetherlands
  - Spain, Austria, Greece...
  - Sweden, Switzerland...
- Extendable to other partners
  - AS (Czech rep., Romania, Bulgaria, Poland, Hungary…)
  - USA, Japan, South Africa, Canada…
- Subcontract to third parties (or partnership)
  - ELRA/ELDA...

European Reseach Area in Language Technologies

- **WP.1. Share of information and best practices**
  - » WP1.1. Programs content
  - » WP1.2. On-going programs in each country
  - » WP1.3. Priorities
  - » WP1.4. Evaluation and Selection process
  - » WP1.5. Management -Best practices in the management of National programs
- **WP2. Strategic activities**
  - » WP2.1. Complementarities
  - » WP2.2. Multinational scheme – Extension of - Participation of other European and Non-European countries
  - » WP2.3. Barriers which prevent from a coordinated European Research Area scheme will be identified
- **WP3. Implementation of joint activities**
  - » WP3.1. Clustering of projects
  - » WP3.2. Joint evaluation of proposals and projects
  - » WP3.3. Testbeds which would agree to welcome the projects of foreign participants will be identified,
  - » WP3.4. Standards
- **WP4. Transnational research activities**
  - » WP4.1. Joint program to enhance multilateral cooperations
  - » WP4.2. Conferences organization
  - » WP4. 3. Preparation of an Article 169 program

ERA-NET

- **III. Partnership**
- A core partnership is made up of 7 founding members having launched national programs in Language Technology.
- *By alphabetical order:*
- Denmark (Ministry of Science, technology and innovation)
- Finland (Academy of Finland)
- France (Ministry of research and new technologies)
- Germany (BMBF)
- Italy (Ministry of Communication)
- The Netherlands (NWO)
- Norway (Ministry of Trade and Industry)
- And a transnational European member: ELRA (the European Language Resource Association) based in Luxembourg. Other comparable entities could be added to fulfill the tasks.
- Other European countries may join very quickly: Austria, Greece, Spain, Switzerland, Sweden, as well as Accessing countries having a long tradition of activities in Language Technology (e.g. Bulgaria, Czech Republic, Hungary, Poland, Romania, Slovenia, etc.) and non European countries having important national programs in Language Technologies: Canada, Korea, Japan, South Africa, USA, or strong activities in that field, and the wish to cooperate such as Israel.
- **IV. Deliverables**
- - Progress and final Reports
- - Transnational experiment (Call; proposals evaluation, choice of participants, results)
- - Conferences and workshop proceedings
- - Web site
- Web structure
- Web portal
- Link to national programs Web sites (should be bilingual: national language + English)
- Language Resources and tools (data, availability, matrix)
- Evaluation (campaigns, participation, results (partly anonymous))
- Standards (link to information)
- Survey (News, reports, studies…)
- Conferences (links to LangTech and LREC Web sites)
- Qualified links (links to major Web sites in the LT area)
- **V. Duration**
- 3 years (2003-2006)
- **VI. Calendar**
- To be discussed for the various tasks.

LREC-2004 Lisbon


GRANADA. Fuente de los Leones en La Alhambra.

WWW.LREC-CONF.ORG

LangTech 2003

d Edition of LangTech (After Berlin 2002

Paris 24-25 November 2003