An Infrastructure for Language Resources sharing and Human Language Technologies Evaluation

Khalid Choukri
ELRA/ELDA

choukri@elda.org
http://www.elra.info/

Overview

- ELRA and Language Resources … Activities,
  - SIGUL (SIG Under Resourced Languages)
- Language Resources (LRs) and Technologies (LTs) through the LREMap
- Public Resources from "Connecting Europe Facilities"
  - *Public Sector Initiative (PSI) Assets & Big Data,*
  - *The new EU General Data Protection Regulation (DGPR)*
- Resources for Regional Languages of France
- Conclusions

The European Language Resources Association (ELRA)

- A not for profit Association of Users of Language Resources for R&D and Technology Development
- Self-sustainable (longetivity) ⇔ Founded in 1995, with strong support from the European Commission
- Activities:
  - Identification and Distribution of LR
  - Production & Validation of LR
  - Technology Evaluation
  - Information Dissemination on Human Language Technologies
- Membership open to
  - European and Non-European Institutions
  - Public and private organizations
- Distribution of 6200+ LR (incl. 1243 LR for free)

The European Language Resources Association (ELRA)

- Operational body: ELDA (European Language Resources Distribution Agency)
  - Incorporated as a company in order to handle all the commercial and business-oriented tasks of the association.
  - Responsible for the development and the execution of ELRA's missions

- ELRA has established the Language Resources and Evaluation Conference (LREC) in 1998 and organized the event ever since.

- Upcoming LREC: 7-12 May 2018 – more on http://www.elra.info/en/lrec/

Language Resources and Evaluation Conference

www.lrec-conf.org





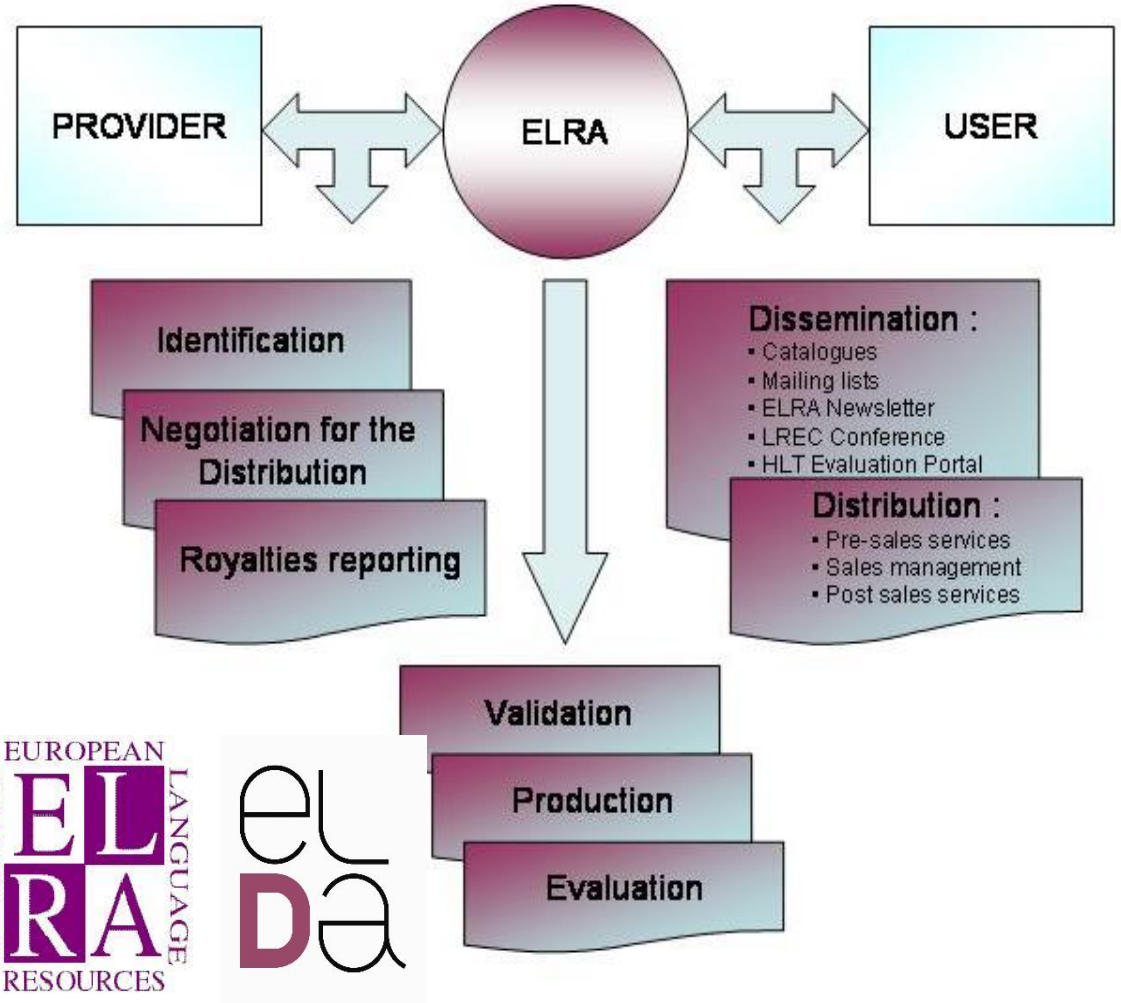
LREC 2018
MIYAZAKI

May 7-12, 2018

Phoenix Seagaia Resort
Miyazaki, Japan



The European Language Resources Association (ELRA)



Membership benefits:

- Discount on LRs prices
- Identification & Production of specific resources
- Universal Catalogue (more info than for general public)
- Legal assistance (Help Desk)
- Evaluation of specific Human language Technologies
- LREC and other events favorable conditions (registration fees, proceedings, …)
- Assistance for validation, production and evaluation activities
- Journal of LR and Evaluation (favorable conditions)
- Fidelity Program
- Market Analysis

ELRA offers / services

Applications

Speech LRs

Automatic Speech Recognition
Speech Synthesis
Speech Translation
Therapy and Speech
Broadcast News Transcription
Acoustic Person Tracking
Acoustic Speaker Identification
Speech Activity Detection

Written LRs

Lexica

Spell Checkers
Information Retrieval
Document Indexing
Information retrieval
Question Answering
Machine Translation
Terminology Extraction
Parsing
Automatic Summarization

Corpora

Database Creation
Lexicon Consolidation
Translation Memory
Validation
Thesaurus Implementation
Generation
Multilingual Text Alignment

Terminology

Multimodal Multimedia LRs

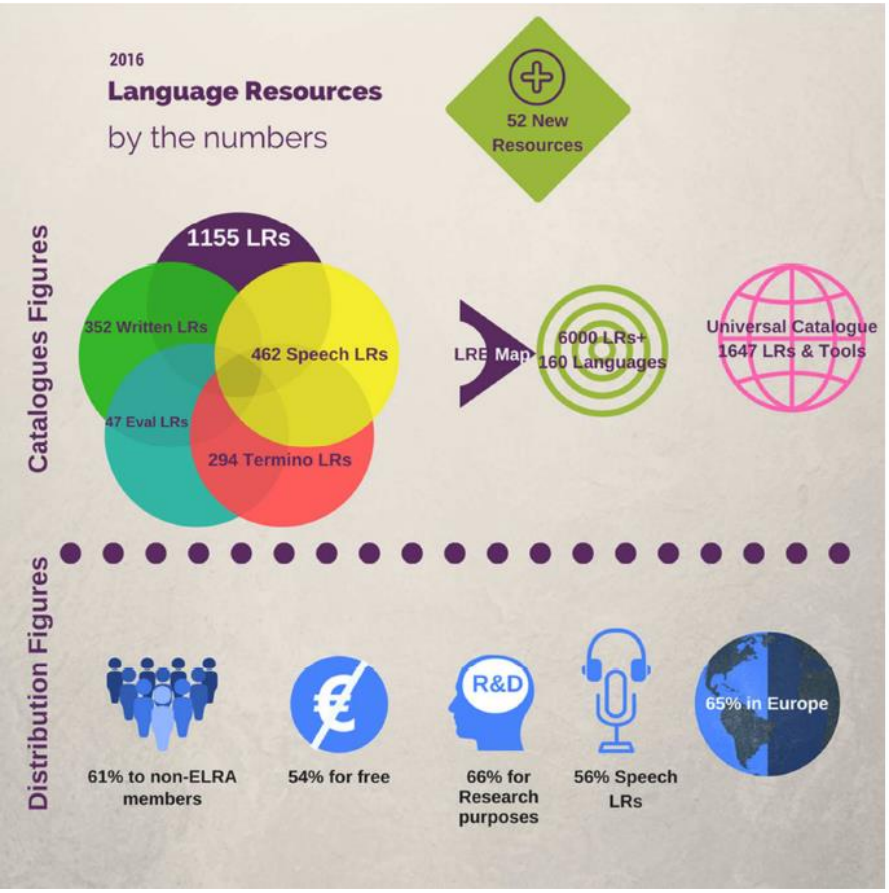
Face tracking
Speech verification
Audio / Mutimodal person identification
Multimodal person tracking
2D head orientation
Audiovisual Speech Recognition

Evaluation Packages / Evaluation Solutions / Evaluation Services

➤➤➤ LRs & Evaluation Packages

(data, metrics, methodologies)

The European Language Resources Association (ELRA)



Identification and distribution of language resources

ELRA Catalogue: http://catalogue.elra.info/

- 1200 Language Resources available off the shelf
- All modalities (text, speech, video, sign, OCR, etc.), numerous languages
- About 30% are free of charge

Universal Catalogue: http://universal.elra.info/

- Information on Language Resources identified world wide
- >2000 identified Language Resources world wide
- Access can be negotiated on-demand

LRE Map: http://lremap.elra.info/

- Information on the use of 7000+ Language Resources and Language Tools
- Mechanism created to monitor the use and creation of Language Resources
- Part of submission process to main NLP conferences (LREC, Interspeech, COLING, etc.)

Some facts about the importance of LRs

- Youtube movies and Internet video growth
  - 400h of video every ….. **Hour**
  - Human Transcriptions … 3-50 times (1h audio = 3h to 50h of labor)
- Translations & Interpretations
  - Over 400.000 translators (150.000 in Europe)
  - Need to translate 506 language pairs in EU, 110 in South Africa, 462 in India, (6000 languages all in all),
  - Not counting converting Sign language to/from Language A
  - Needs for translation grow by 30% …. **every year**
  - Consensus: 10% of data is translated

Management of Bilingual Data Example

Alignement of English and French versions

S1. Executive Summary

S2. This report is the result of a collective work carried out by the high-level expert Committee and a writing team commissioned by the Task Force on Innovative Financing for agriculture, food security and nutrition created by the Leading Group on Innovative Financing for Development at its 9th plenary session in Mali (Bamako) in June 2011.

S3. The report includes an analysis of the need for innovating financing dedicated to the agricultural, food security and nutrition sector, a critical review of existing and possible mechanisms and a proposed selection of avenues for the development of such mechanisms on the basis of the expertise of a high-level Committee of experts, literature review, meetings with relevant professional actors and an on-line consultation on the Global Forum on food security and nutrition (FSN Forum)1.

S4. The setting up of the Task Force on Innovative Financing for agriculture, food security and nutrition responds to current and future crucial challenges faced by the international community [...]

S1. Résumé

S2. Le présent rapport résulte d'un travail collectif mené par le Comité d'experts de haut niveau et une équipe de rédacteurs désignés à cette fin par le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition.

S3. Ce groupe de travail a été créé par le Groupe pilote sur les financements innovants pour le développement lors de sa 9e session plénière, qui s'est tenue au Mali (Bamako) en juin 2011.

S4. Le présent rapport comporte une analyse des raisons pour lesquelles des financements innovants dédiés à l'agriculture, à la sécurité alimentaire et à la nutrition sont nécessaires, propose un examen critique des mécanismes existants et possibles, et présente une sélection de méthodes pour mettre au point ces mécanismes.

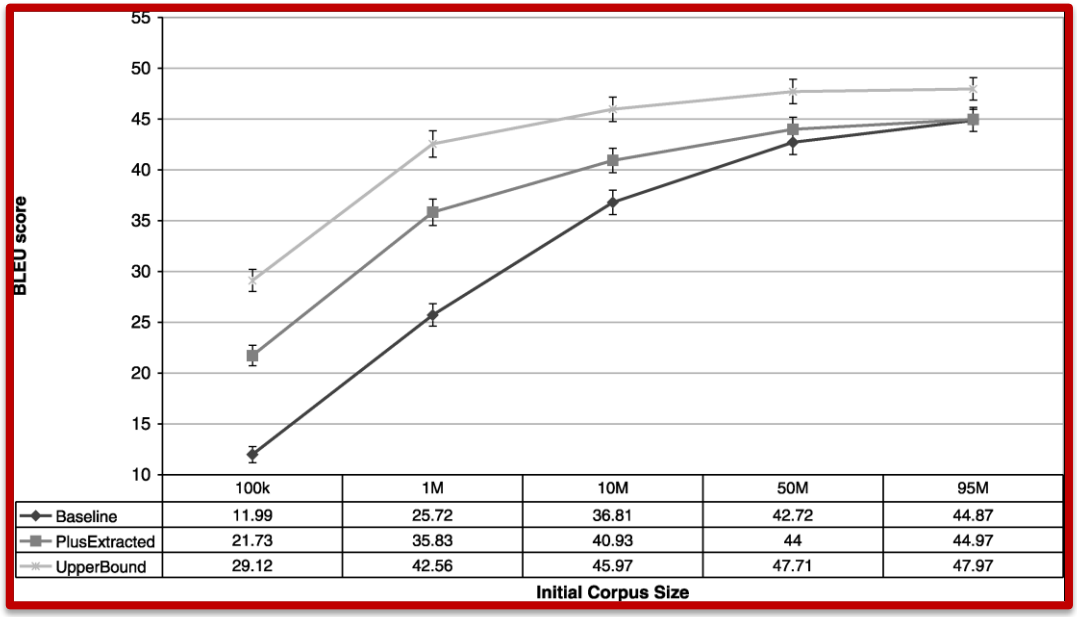
S5. Il s'appuie à ces fins sur l'expertise du Comité d'experts de haut niveau, une analyse bibliographique, des réunions avec les professionnels concernés et la consultation en ligne organisée par le Forum global sur la sécurité alimentaire et la nutrition (Forum FSN)1. ...]

Importance of data and Re-usability

- ✓ **Almost all technologies are data driven and based on statistical paradigms …**

(modeling based on huge amounts of date)

Let us look at MT performance when "simply" adding data



	100k	1M	10M	50M	95M
Baseline	11.99	25.72	36.81	42.72	44.87
PlusExtracted	21.73	35.83	40.93	44	44.97
UpperBound	29.12	42.56	45.97	47.71	47.97

MT performance improvements for Arabic-English

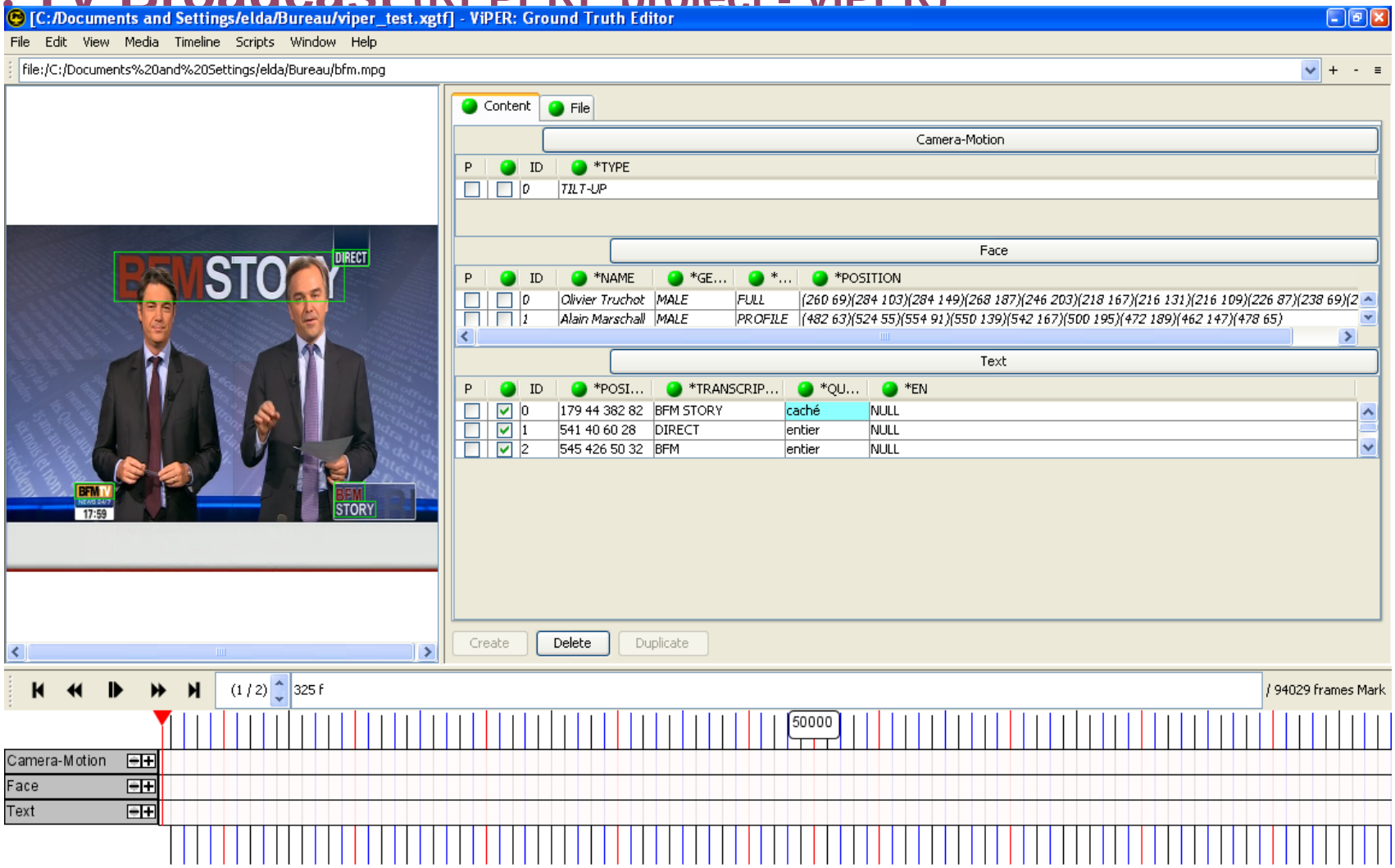
(Courtesy Dragos Stefan Munteanu and Daniel Marcu)

Collecting/Producing data is costly … time … money …

Archiving and sharing is essential

Video Annotation

TV Broadcast (REPERE project - ViPER)



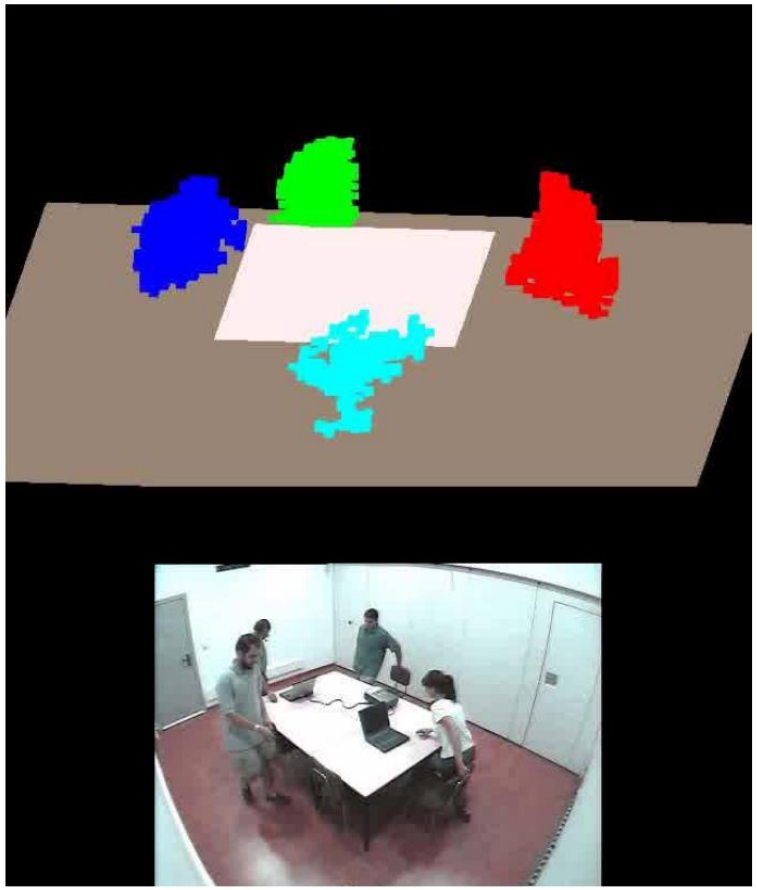
Video Annotations

- **Meeting room captured by 5 cameras (CHIL project)**





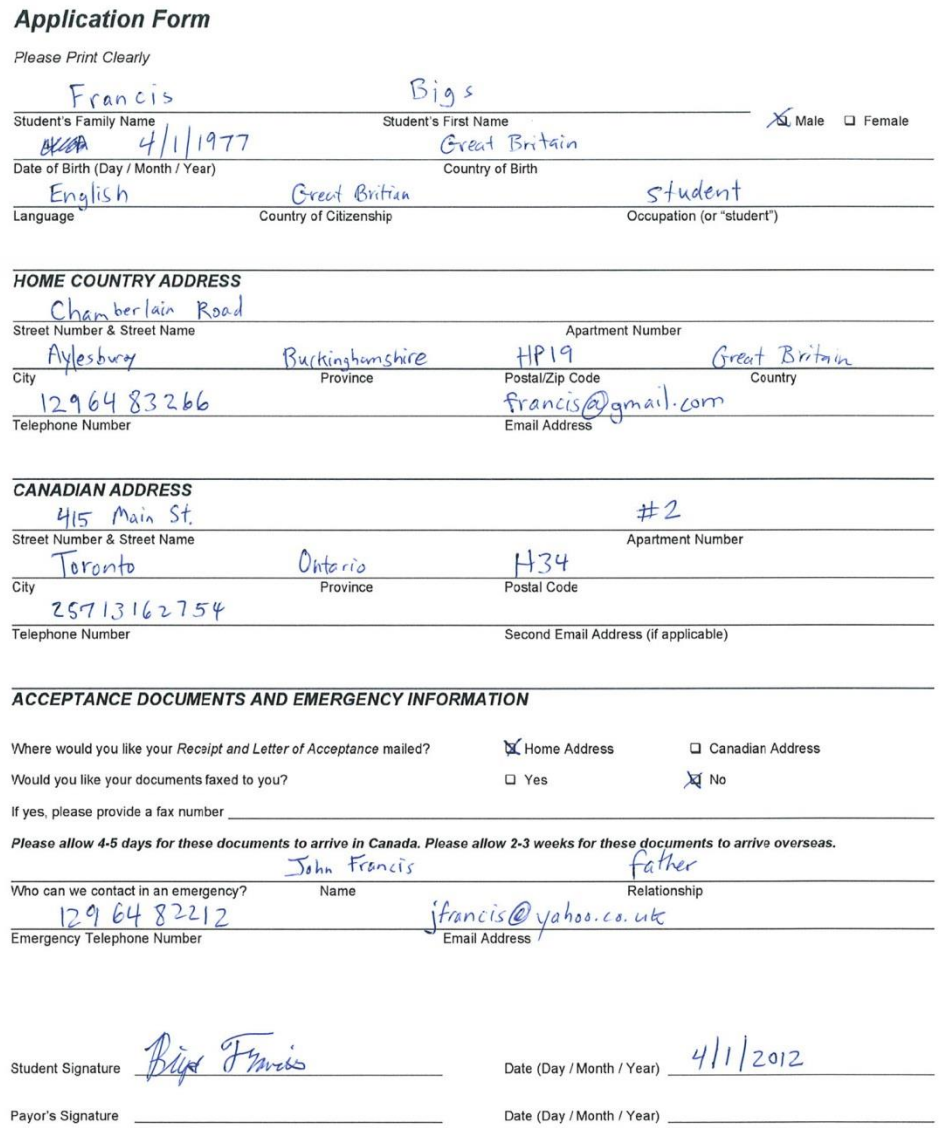
Information fusion from multiple sensors and joint detection and tracking in 3D (UPC)



Evaluation & Evaluation-Oriented LR Production

- **Multimodal Technologies …** MAURDOR Project
  - Automatic processing of digital documents
  - Objective: improve processing technologies for handwritten and typewritten documents in AR, EN, FR
  - Large production and annotation of necessary documents for evaluation: 10K documents
  - Hardcopy scans: representative of daily human operations
  - To avoid IPR infringement: large part created from scratch + negotiated
  - Very specific regarding languages, handwritten/typed, number of contributors/scribes

Examples …



Application Form

Please Print Clearly

Francis — Student's Family Name
Bigs — Student's First Name
☒ Male ☐ Female
4/1/1977 — Date of Birth (Day / Month / Year)
Great Britain — Country of Birth
English — Language
Great Britain — Country of Citizenship
student — Occupation (or "student")

HOME COUNTRY ADDRESS

Chamberlain Road — Street Number & Street Name
Apartment Number
Aylesbury — City
Buckinghamshire — Province
HP19 — Postal/Zip Code
Great Britain — Country
12964 83266 — Telephone Number
francis@gmail.com — Email Address

CANADIAN ADDRESS

415 Main St. — Street Number & Street Name
#2 — Apartment Number
Toronto — City
Ontario — Province
H34 — Postal Code
2571316 2754 — Telephone Number
Second Email Address (if applicable)

ACCEPTANCE DOCUMENTS AND EMERGENCY INFORMATION

Where would you like your *Receipt and Letter of Acceptance* mailed? ☒ Home Address ☐ Canadian Address

Would you like your documents faxed to you? ☐ Yes ☒ No

If yes, please provide a fax number

Please allow 4-5 days for these documents to arrive in Canada. Please allow 2-3 weeks for these documents to arrive overseas.

John Francis — Who can we contact in an emergency? Name
father — Relationship
12964 82212 — Emergency Telephone Number
jfrancis@yahoo.co.uk — Email Address

Student Signature — Date (Day / Month / Year) 4/1/2012

Payor's Signature — Date (Day / Month / Year)

Categories


مسجد القدس
البرج
فرع المصرف
التجاري السوري
شارع عبد الله بن عبد الله
شارع السينما
سينما
الأهرام
صرف
التجاري

Document analyses and

HARTEL Michael
2 rue du Tilleul
68480 Fislis
Tél : 0372162771

Objet : Modification de co

Madame, Monsieur,
Je vous contacte au
commande de CD vier
le 03/05/2006. Ma r
MYRN947. Je sou
quantité des CDs

Cordialement,

Special Interest Group: Under-resourced Languages (SIG-UL)

Courtesy of Sakriani Sakti

Slides prepared by the SIGUL Secretary: Sakriani Sakti (NAIST, Nara, Japan)

SIG UNDER-RESOURCED LANGUAGES

- **Joint Special Interest Group:**
  - ✓ International Speech Communication Association (ISCA)
  - ✓ European Language Resources Association (ELRA)
  - ✓ (Previous SIG-SaLTMil members have approved the creation of SIGUL)
- **Main objectives:**
  - ✓ Supports linguistic diversity through technology and ICT
  - ✓ Commits to increasing the lesser-resourced languages (regional, minority, or endangered) chances to survive in the digital world
- **Board (provisional before elections):**
  - ✓ Chair and ELRA liaison representative: Claudia Soria (CNR-ILC, Pisa, Italy)
  - ✓ Co-chair and ISCA liaison representative: Laurent Besacier (LIG, France)
  - ✓ Secretary: Sakriani Sakti (NAIST, Nara, Japan)

SIG Supported Activities

- **Workshops:** SLTU/CCURL every 2 years

  - ✓ **"Collaboration and Computing for Under-Resourced Languages Towards an Alliance for Digital Language Diversity (CCURL)"**

    Submitted as LREC Satellite Workshop (waiting for acceptance)

    Venue: 7-8 May or 12 May 2018 (Miyazaki, Japan)

  - ✓ **"Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)"**

    Interspeech 2018 Satellite Workshop

    Venue: 30-31 August 2018 (New Delhi, India)

SIG Supported Activities

- **Tutorials**

  Speech and language technologies for UL communities

- **Panel Meeting/Discussions**

  Between various institutions

- **Language Resources**

  Help to monitor the use and creation of under-resourced language resources and technologies (LREMap & Language Matrices)

SIG-UL Members

- **Supported Members:**
  - ✓ Previous SLTU+SaLTMiL+CCURL members that have confirmed to join SIGUL
  - ✓ Total: 201 Members from 34 Countries
  - ✓ Currently preparing a *census* for future elections
- **Join SIG-UL**
  - ✓ If you are interested in joining SIG-UL please send email inquiries to

    sigul-board@list.elra.info
- **More Details:**
  - ✓ ISCA: http://www.isca-speech.org/iscaweb/index.php/sigs
  - ✓ ELRA: http://www.elra.info/en/sig/sigul/

ELRA offers / services

Applications

Speech LRs

Automatic Speech Recognition
Speech Synthesis
Speech Translation
Therapy and Speech
Broadcast News Transcription
Acoustic Person Tracking
Acoustic Speaker Identification
Speech Activity Detection

Written LRs

Lexica

Spell Checkers
Information Retrieval
Document Indexing
Information retrieval
Question Answering
Machine Translation
Terminology Extraction
Parsing
Automatic Summarization

Corpora

Database Creation
Lexicon Consolidation
Translation Memory
Validation
Thesaurus Implementation
Generation
Multilingual Text Alignment

Terminology

Multimodal Multimedia LRs

Face tracking
Speech verification
Audio / Mutimodal person identification
Multimodal person tracking
2D head orientation
Audiovisual Speech Recognition

Evaluation Packages / Evaluation Solutions / Evaluation Services

➤➤➤ LRs & Evaluation Packages

(data, metrics, methodologies)

- DO WE HAVE THESE TECHNOLOGIES AND THE ASSOCIATED RESOURCES FOR ALL LANGUAGES ? (or at least the basics RESOURCES?)

- The language digital extinction or the BLARK/ELARK Revisited (Basic/Extended Language Resource Kits)

LRE Map

(http://lremap.elra.info)

With contribution from Joseph Mariani

LRE Map : a LR analysis instruments

- The concept: collect and compile data from all papers of LREC, (initiated @LREC 2010)

  N. Calzolari, C. Soria, R. Del Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani, S. Piperidis, The LREC 2010 Resource Map, LREC'2010, Malta, 19-21 may 2010

- **Associate LRs and Publications** /metadata descriptions
- Data supplied by the authors (experts) in a structured way

  - ✓ Provide state of the art: Clear picture of what exist per language
  - ✓ Identify Gaps and missing kits (LRE-Map Matrices),
  - ✓ Design & set in motion clear roadmaps

Exploitation of the LRE'MAP to refine BLARK data through Language Resource Matrices

Conferences in the LRE Map

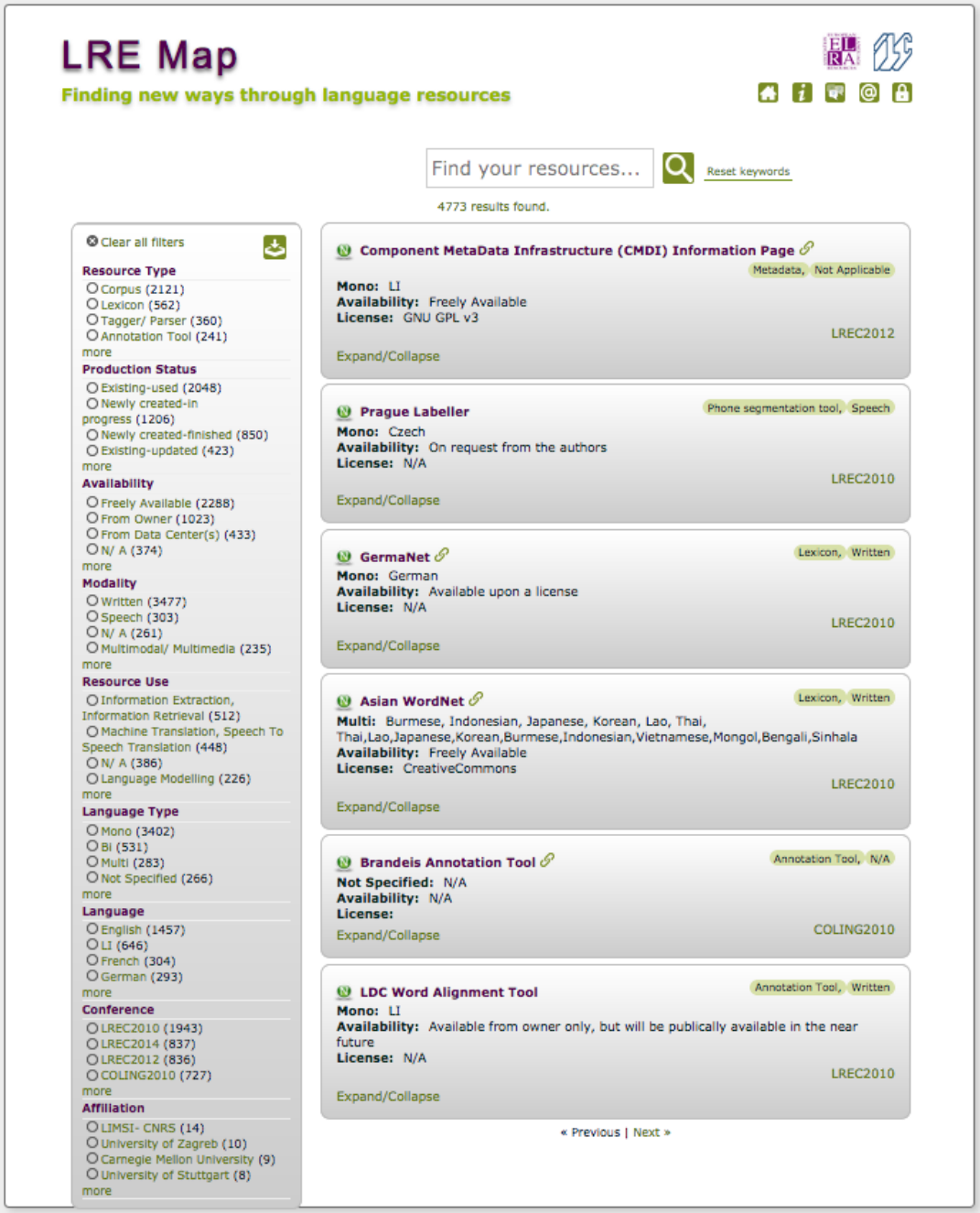
- LREC2010
- COLING2010
- ACLHLT2011
- IJCNLP2011
- Interspeech2011
- LTC2011
- RANLP2011
- O-COCOSDA2011
- LREC2012
- COLING2012
- NAACL2013
- Interspeech2013
- RANLP2013
- LREC2014, COLING
- LREC2016 (inc. satellite workshops) , COLING

~~7000 LR entries

LRE Map metadata

- Contains list of LR with related attributes
  - Paper id (conference, year, Paper id, LR number)
  - LR Name
  - Type, e.g. Corpus, Lexicon, Annotation Tool, Tagger/Parser
  - Modality(ies), e.g. Speech, Written, Sign language
  - Language(s), e.g. English, Estonian, Multilingual, L.I.
  - Size, e.g. 50 Gbytes, 20 hours, 50 Mwords
  - Production Status, e.g. Newly created, Completed
  - Use, e.g. Speech recognition, Machine Translation
  - Availability, e.g. Freely available, From Data Center
  - Licensing, e.g. Creative Commons, LDC
  - Documentation
  - URL

http://lremap.elra.info



LR Matrices

- LR Matrices developed within T4ME/META-NET EC project. (J. Mariani et. al.)
- Objective of the LR Matrices:
  - Provide a clear picture of what exists in terms of Language Resources (LR) across languages
    - Data, Tools, Evaluation, Meta-Resources (Standards, Metadata, Guidelines)
  - Compile facts & figures about existing LR, and access
  - Identify the language gaps in order to fill them

J. Mariani, G. Francopoulo, Language Matrices & the Language Resource Impact Factor, in Language Production, Cognition, and the Lexicon, N. Gala, R. Rapp, G. Bel eds., Springer, December 2014

Revisiting BLARK through the LRE Map
The Language Matrices 1/

- Automatically built from the LRE Map
  - Matrix Generator Software by G. Francopoulo
- Various LR Matrices
  - All types of LR
  - Written Language Data and Tools
  - Spoken Language Data and Tools
  - Multimodal/Multimedia Data and Tools
  - Evaluation
  - Meta-Resources
  - Specific Languages,…

Language Matrix … Study language coverage
Spoken Language Data

Typology of LR

Languages

	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish	Other Europe	Asturian	Basque	Catalan	Galician	Arabic	Hindi	Japanese	Korean	Mandarin	Other	Multilingual	L.I.	N.A.	Total
Corpus	1	4	7	12	80	1	1	35	22	2	2	0	9	0	0	0	7	7	1	0	1	18	5	26	0	6	4	3	15	1	12	3	11	31	2	4	3	336
Lexicon	0	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	11	0	0	0	18
Gram-mar/Language Model	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
Acoustic and language models	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Ontology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Statistical Speech Resource	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Total	1	5	7	12	87	1	1	35	25	2	2	0	9	0	0	0	7	7	1	0	1	19	6	26	0	7	4	3	16	1	12	3	11	42	2	5	3	363

English : 87 items

Spanish: 19 items

Arabic: 16!

LRE Map

- **LRE Map**
  - **Major source to revisit the BLARK concept /Language**
  - **Normalisation of LR names** and other values
  - Through auto-completion , offering the normalised name, we push for normalisation
  - LRE-Map **as a regular feature for other NLP Conferences?**
  - **Impact Factor …**
- **LRE Map an instrument for LR Discovery and Road mapping only if heavily used and representative of the community assets**
- **Please use it , promote it**

The European Language Resource Coordination (ELRC) Initiative "Connecting Europe Facility (CEF)"

EU Action 3: Open up public data resources for re-use

- Public authorities produce large amounts of data that could become the raw material for new, innovative cross-border applications and services. Examples of products and services based on the re-use of Public Sector Information (PSI) are GPS, weather forecasts, financial and insurance services.

- **What about Data expressed in Words !!**

The issue

- Need for:
  - a clear and easy to follow regime for data re-use across the EU
  - legal and technical interoperability
  - simple redress mechanisms
- Objective:
  - To develop a single European market for innovative apps based on public data
- Privileged Framework
  - Public Sector Initiative (PSI)
  - Directive 2003/98/EC / revised by Directive 2013/37/EU
- French context
  - Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal
  - Décret n° 2011-577 du 26 mai 2011 relatif à la réutilisation des informations publiques détenues par l'Etat et ses établissements publics administratifs

European Language Resource Coordination
Connecting Europe Facility

What has been achieved so far…?

Collection of 225 language resources overall

More than 2 billion words in all EU official languages, Norwegian and Icelandic

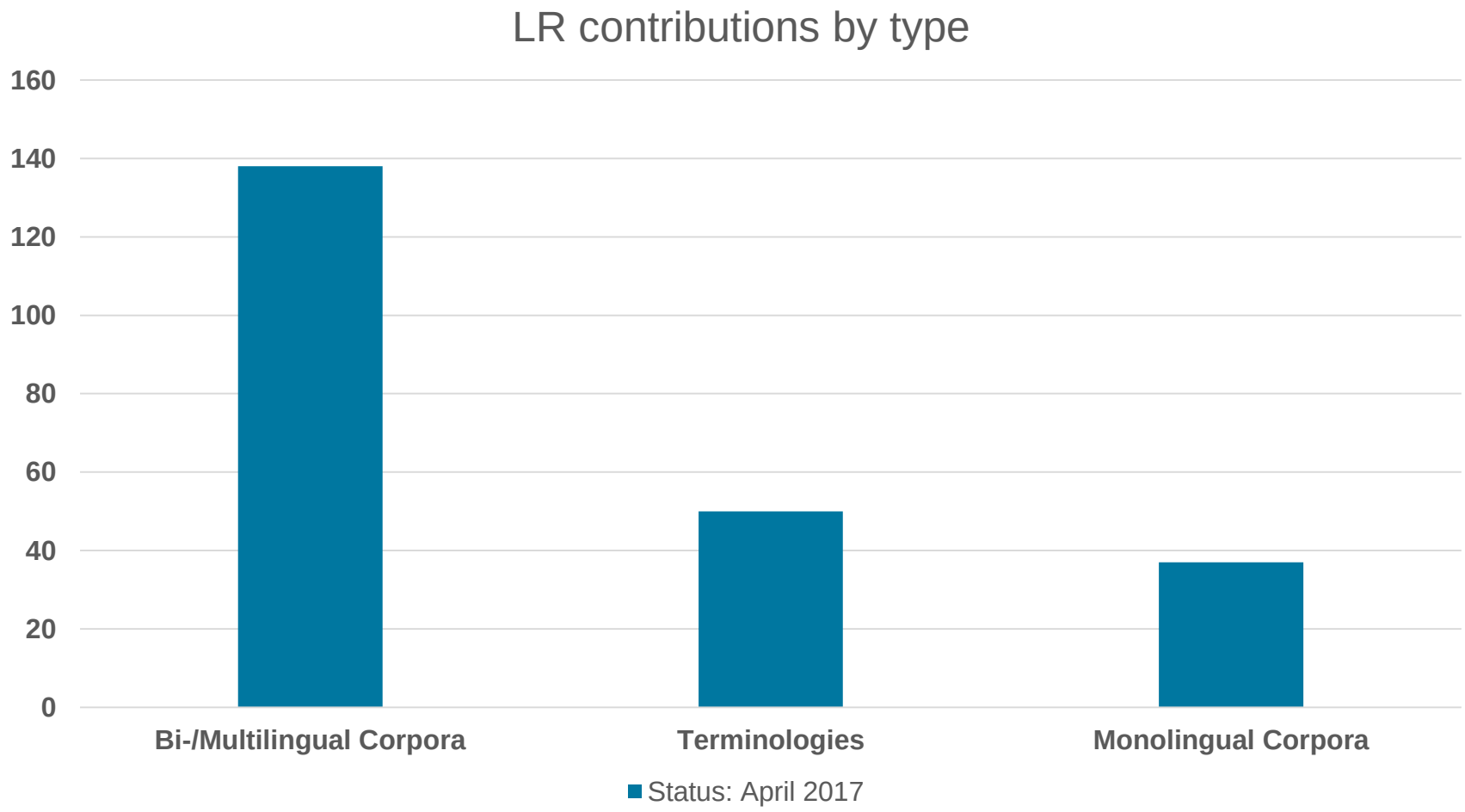
More than 91 language resources to be used by you!

Over 450.000 terms

More than 2 million translation units

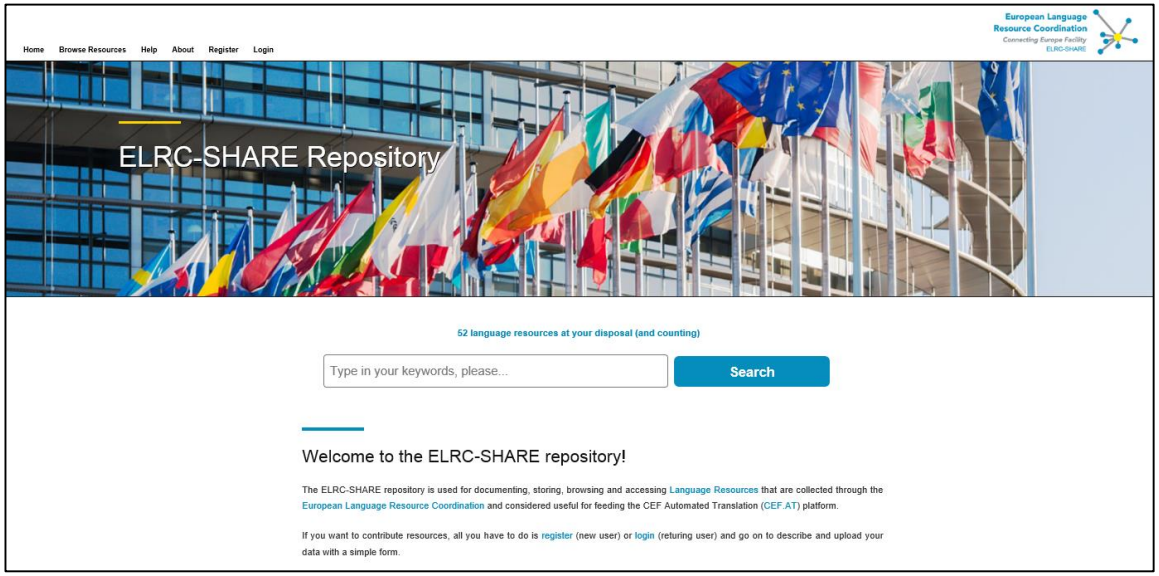
European Language Resource Coordination
Connecting Europe Facility

What has been achieved so far…?



ELRC for you…

- **The ELRC-SHARE Repository**
  - Access to, sharing and contribution of language resources
  - Access to tools and services catalogue (upcoming)
  - Visit http:/www.lr-coordination.eu/resources



GDPR … Yet another directive

- they significantly reinforce the protection of individuals against unlawful processing of their personal data and -- correlatively -- the obligations of data controllers. The most important changes include:
- the principle of accountability, according to which it is the data controller that has to be able to demonstrate compliance with the GDPR;
- the obligation to implement "**privacy by design and by default**";
- ***the obligation to carry out -- in certain cases relevant for the language resources community -- a data protection impact assessment;***
- the obligation (for certain categories of data controllers) to appoint a data protection officer;
- the obligation to implement appropriate safeguards (such as pseudonymisation or encryption) for processing of personal data for research purposes;
- the obligation to notify data breaches to the supervisory authority and to data subjects;
- the right to data portability;
- reinforced **transparency of processing, including in particular the use of algorithms on personal data;**
- ……

- Infringements of the GDPR can be subject to fines up to 20 000 000 EUR, or up to 4% of the total worldwide annual turnover of the preceding financial year.
-

Review on the Existing Language Resources for Languages of France

- Partnership with the *Délégation générale à la langue française et aux langues de France* (DGLFLF, French Ministry of Culture and Communication)
- Two projects 2013-2014, 2017-2018
- Main goal: assess exploitability of identified resources within different kinds of language technologies.
  - more visibility and applicability among a wider audience for regional languages with adapted technology
- Tasks:
  - identify main channels of production and dissemination (sources)
  - provide a non-exhaustive list of existing LRs.

Review on the Existing Language Resources for Languages of France

- 84 languages spoken in France and its overseas departments and territories
- LR Metadata inspired from existing nomenclatures (ELDA, LDC, OLAC):
  - LR type (written corpus, spoken corpus, parallel corpus, multimedia resource, lexicon, grammar, thesaurus)
  - Name & Description
  - Related language(s), classified with different language families, number of speakers, transmission modalities (oral, written or signed)
  - Volume
  - Applications that can be developed with the LR
  - Location on the Internet,
  - Provider(s)
  - Availability and possible rights associated to LR usage

Review on the Existing Language Resources for Languages of France

- Collection in a MySQL database, hosted at ELDA
- Inclusion of Metadata

- 2,299 LRs identified:
  - 1,417 spoken corpora
  - 425 written corpora, including parallel corpora
  - 181 lexica
  - 206 multimedia/multimodal corpora
  - 16 grammars/language models
  - 1 ontology
  - 7 thesauri/wordnets
  - 17 media (newspapers) collections
  - 19 TV/Radio resources
  - 10 mixed corpora, i.e. combining several types of LRs

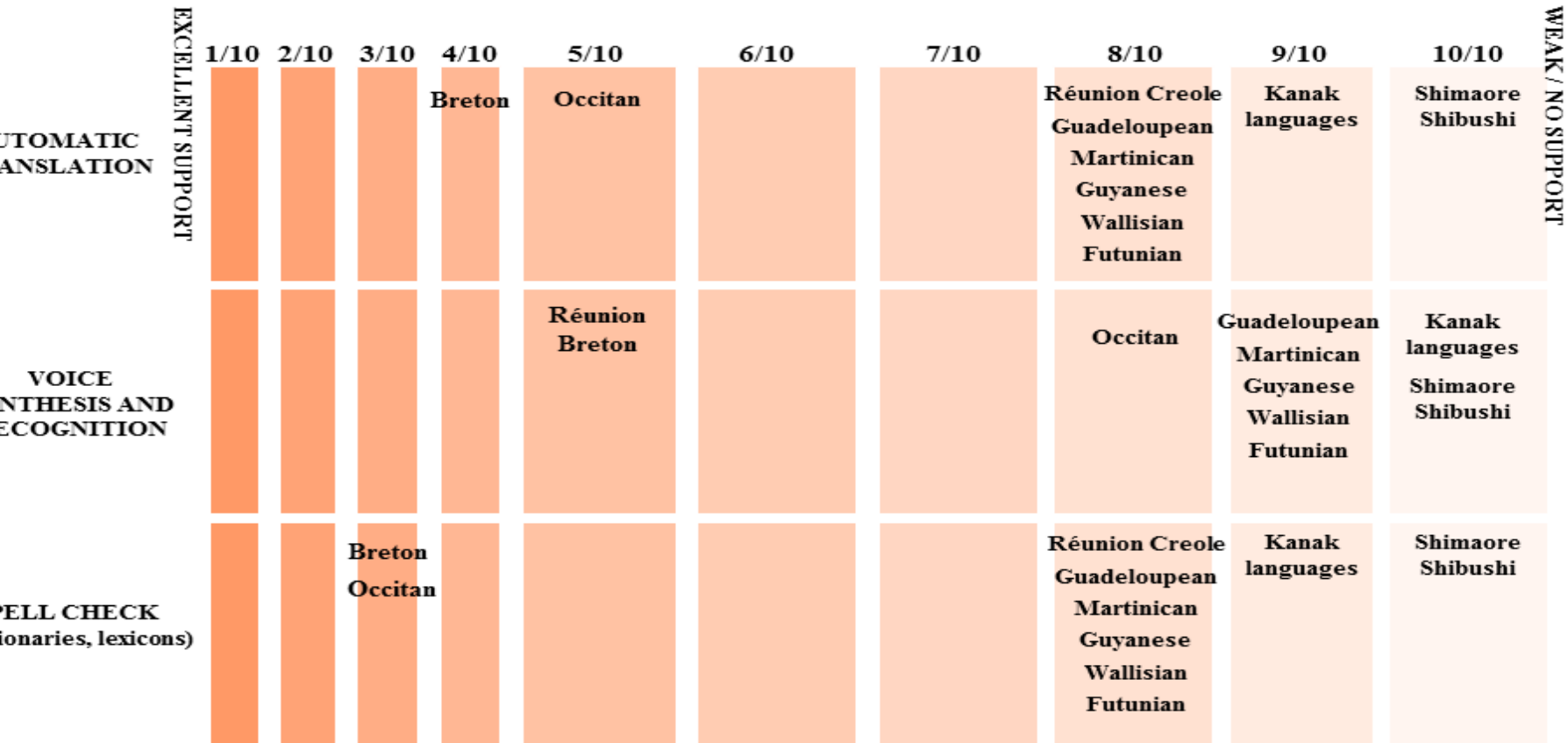
Language	Nb	Rank
Ajië	3	46
Allemand	4	43
Alsacien	127	6
Aluku	7	38
Anglais	11	34
Arabe Dialectal	29	17
Arménien Occidental	1	62
Auvergnat	16	27
Basque	17	24
Breton	420	1
Catalan	47	10
Cèmuhi	9	37
Corse	93	7
Drehu	14	31
Espagnol	9	36
Fagauvea	26	19
Flamand Occidental	2	54
Français	10	35
Francoprovençal	2	47
Futunien	44	12

Language	Nb	Rank
Galicien	1	61
Gascon	286	2
Grec moderne	2	50
Guadeloupéen	14	30
Guyanais	4	45
Iaai	13	32
Italien	4	44
Judéo-Espagnol	164	5
Kumak	70	8
Langue des Signes Allemande	2	51
Langue des Signes Britannique	2	53
Langue des Signes Française	47	11
Langue des Signes Grecque	2	49
Languedocien	202	3
Latin	1	63
Limousin	27	18
Mahorais (shimaore)	4	42
Marquisien	191	4
Néerlandais	1	58
Nemi	31	15
Nengone	2	52

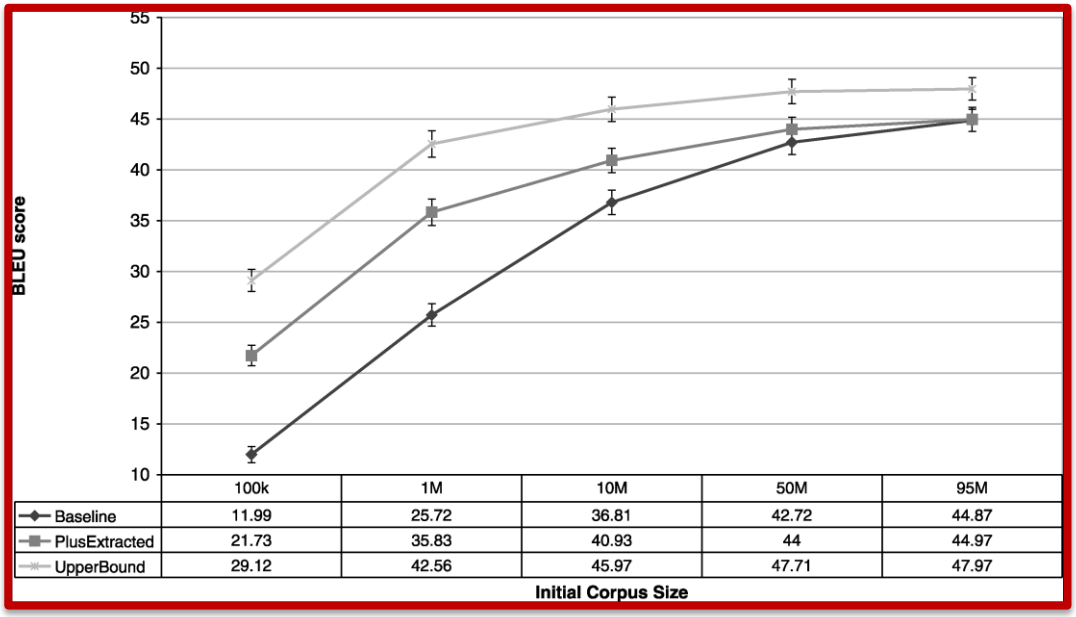
Language
Niçois
Numèè
Occitan
Paicî
Palikur
Picard
Pije
Portugais
Provençal
Réunionnais
Romani
Roumain
Suédois
Vivaro-Alpin
Wallisien
Wallon
Wayampi
Wayana
Xâracùù
Xaragurè
Yuaga

Focus

Languages	SPEECH CORPORA	WRITTEN CORPORA	MULTIMÉDIA CORPORA	PARALLÈL CORPORA	LEXICA	THÉSAURUS	OTHERS
Breton	403	17					
Occitan	365	92	29	33	102	1	
Overseas	300	159	106		1		20



LRs and Big Data



Zero data training

Neural networks / deep learning

Thank you for your attention