Deliverable D3.7

Coordination with and support of ICT-17a and ICT-17b projects

Authors:	Kanella Pouli (ATHENA RC)
	Stelios Piperidis (ATHENA RC)
Dissemination Level:	Public
Date:	31 December 2017

Grant agreement no.	645357
Project acronym	CRACKER
Project full title	Cracking the Language Barrier
Type of action	Coordination and Support Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	1 January 2015, 36 months
Dissemination level	Confidential
Contractual date of delivery	31.12.2017
Actual date of delivery	31.12.2017
Deliverable number	D3.7
Deliverable title	Coordination with and support of ICT-17a and ICT-17b projects
Type	Report
Status and version	Final
Number of pages	15
Contributing partners	DFKI
WP leader	ATHENA RC
Task leader	ATHENA RC
Authors	Kanella Pouli (ATH), Stelios Piperidis (ATH)
Internal reviewer(s)	Georg Rehm (DFKI)
EC project officer	Pierre-Paul Sondag (M01-M18), Susan Fraser (M19-M36)
The partners in CRACKER are:	• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany • Charles University in Prague (CUNI), Czech Republic • Evaluations and Language Resources Distribution Agency (ELDA), France • Fondazione Bruno Kessler (FBK), Italy • Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece • University of Edinburgh (UEDIN), UK • University of Sheffield (USFD), UK

For copies of reports, updates on project activities, and other CRACKER-related information, contact:

DFKI GmbH
CRACKER
Dr. Georg Rehm   georg.rehm@dfki.de
Alt-Moabit 91c   Phone: +49 (0)30 23895-1833
D-10559 Berlin, Germany   Fax: +49 (0)30 23895-1810

Copies of reports and other material can also be accessed via http://cracker-project.eu/.

Contents

1   Introduction

The objective of the CRACKER project, as described in the Description of Action (DoA), has been to improve Machine Translation (MT) research in various dimensions, one of them being language resources sharing as well as coordination and support of ICT17a and b projects. In order to do so, CRACKER has used as a nucleus MT-related projects funded through H2020-ICT-17a/b, extending also to relevant FP7 actions such as QTLeap, LIDER and MLi, as well as related initiatives. CRACKER's aspiration has been to create for the MT research community a bridge between the emerging challenges and the available language resources which would eventually result in better MT results in terms of efficiency and effectiveness.

This Deliverable D3.7 gives an overview of the coordination and support of these projects which was achieved through a series of instruments and actions, including:

a) the setup of the Cracking the Language Barrier federation, the first members of which have been the ICT17a, b and c projects,

b) the Strategic Research and Innovation Agenda (SRIA),

c) guidelines and recommendations for appropriate data management promoting language resource sharing,

d) joint communication and dissemination activities.

2   Cracking the Language Barrier federation

CRACKER initiated the Cracking the Language Barrier[1] federation. The purpose of the federation has been the creation of a strong community of MT-related projects funded by the ICT-17-2015 call of Horizon 2020. CRACKER has assembled and continues to attract projects and organisations working with multilingual technologies in various language technology fields and applications.

To be able to do so a coordination board was established[2] the responsibilities of which spanned from "streamlining communication among its members and the EU-projects they represent" to "stimulating cooperation and collaboration, and identifying synergies, common topics and goals etc." The members of the board, representing the respective projects, have been given the mandate "to take strategic high-level decisions that concern the whole setup of ICT-17-projects" and represent them in various events, conferences and industry meetings, for example. The kick-off meeting[3] of the ICT-17-projects (which also included MLi and QTLeap) took place on April 28th, 2015 in Riga, Latvia, in the framework of META-FORUM 2015 and the Riga Summit 2015.

---

[1] "Cracking the Language Barrier" is the name adopted for the described "QT Initiative" in the Description of Action.

[2] The Cracking the Language Barrier board corresponds to the "QT Board" described in the Description of Action.

[3] The kick-off meeting is reported in CRACKER Deliverable D1.2, "Kick-off meeting of the ICT-17 group of funded projects".

A brand image was adopted but not a separate initiative brand, allowing for each of the projects participating to maintain their own identity manifested also by their brands. The logo chosen (Figure 1) depicts the European approach and scope which is also the common denominator of all projects.

Figure 1: Cracking the Language Barrier federation logo

Discussions throughout the project led to the decision to extend the federation scope so as to include organisations[4], the main rationale being that organisations have a longer lifespan compared to projects which are of limited duration. This would provide to the federation a more stable and sustainable basis in the future.

2.1 Memorandum of Understanding

Interested parties wishing to join the federation have to sign a Memorandum of Understanding (MoU) (see Deliverable D1.7). All parties participating in the kick-off meeting agreed that the MoU should be a short and clear document with no legal restrictions imposed on those signing it[5]. In fact the MoU expresses the federation's desire to establish and reinforce a joint effort for projects sharing a common objective, "cracking the language barrier" and similar goals related to MT enhancement; it also outlines the ways to create a community of sharing resources and ideas. This initiative is also expected to have a stronger longer lasting voice in upcoming EU discussions as concerns policies addressing issues of multilingualism, such as the Digital Single Market.

The MoU also foresaw the creation of a dedicated website[6], distinct from the project's website, as well as new communication channels (e.g., mailing lists) to serve the federation's purposes. The website provides the space where all projects and organisations are featured, each having their dedicated page with a brief description. The website is subdivided to an informative section (About), one page per project or organisation, a section dedicated to the project results (Deliverables and Documents) and a section for events (where information on past and upcoming events can be found).

---

[4] Such is the case of the organisation CITIA (created by the ROCKIT project).

[5] Minimum participation entailed representation both in the federation's website and mailing lists. The MoU was circulated via email to all ICT-projects on May 26th, 2015. After a short round of discussions minor changes were made and it reached its final form.

[6] The website was launched in late August 2015 under the domain cracking-the-language-barrier.eu. A webserver was set up using the most recent version of WordPress on a machine that responds to the URL http://www.cracking-the-language-barrier.eu. DFKI will administer and maintain the federation's website and mailing lists for the period 2018+, i.e., after the end of the funding period for CRACKER.

Figure 2: The landing page of the Cracking the Language Barrier website

The mailing lists[7] serve so as to inform participants on issues of general or specific interest (e.g., workshops, conferences, etc.). Depending on the desired level of engagement participants can also contribute with website content (i.e., blog posts, event announcements) and participation in several activities (workshops for instance).

At the time of writing (December 2017) the MoU has been signed by 25 projects[8] and 12 organisations[9] while the federation is still in the process of attracting new candidates.

As decided at the kick-off meeting, and depicted in the MoU, the federation came up with an action plan which was carried out with the help of all members. The actions foreseen included:

- strategic work such as, e.g., the Strategic Research and Innovation Agenda (SRIA), mainly driven by CRACKER and LT_Observatory,

- a Data Management Plan template, including suggestions for repositories where datasets can be deposited, legal aspects of data sharing, metadata for data documentation, etc. (mainly driven by CRACKER, making use of the principles and tools provided by META-SHARE)

---

[7] The following lists were set up: projects@cracking-the-language-barrier.eu (for the project coordinators), contact@cracking-the-language-barrier.eu (for the respective contact persons) and office@cracking-the-language-barrier.eu (for contacting the office of the federation).
[8] For a full list of projects having signed the MoU see CRACKER Deliverable D1.7, "Final Report on Coordination among QT Projects".
[9] For a full list of organisations having signed the MoU see CRACKER Deliverable D1.7, "Final Report on Coordination among QT Projects".

- organisation and participation in external communication, events (such as, for example, META-FORUM or workshops), a communication plan and branding (mainly driven by CRACKER)

- shared tasks and evaluations (mainly driven by QT21)

3 Strategic Research and Innovation Agenda

The Strategic Research and Innovation Agenda (SRIA) has been one of the most important issues addressed by the federation. It is collaborative work, which represents the current state of discussion within the language technology research, development and innovation community touching upon the status quo and key challenges raised, applications and solutions (such as e-learning, e-health, etc.), research themes (for example, conversational technologies and high quality machine translation), the horizontal topics of interoperability and copyright and data protection among others before reaching at the expected economic impact and potential funding sources.

The preparation of the Strategic Research and Innovation Agenda has so far resulted in three versions. All are accessible on the website:

- Strategic Agenda for the Multilingual Digital Single Market (SRIA V0.5, published in April 2015)[10]

- Strategic Research and Innovation Agenda for the Multilingual Digital Single Market (SRIA V0.9, published in July 2016)[11]

- Language Technologies for Multilingual Europe: Towards a Human Language Project (SRIA V1.0, published in December 2017)[12]

While the SRIA Version 0.5 was prepared by CRACKER in collaboration with the second Coordination and Support Action funded through ICT-17, LT_Observatory, the other two versions of the SRIA (V0.9 and V1.0) were prepared in collaboration with the members of the Cracking the Language Barrier federation, which also includes LT_Observatory as a member.

3.1 The Future

The federation is based upon a very agile and lean structure and does not have a governance structure. Future developments as concerns Multilingual Europe may impose changes to which the federation will adjust, if necessary. The federation will in the meantime continue to disseminate its message to the community and attract new members.

---

[10] Deliverable D5.5: "Position Paper and preliminary joint SRIA for the LT/MT field".
[11] Deliverable D5.6: "SRIA for the Language Technology/Machine Translation field.
[12] Deliverable D5.7: "SRIA for the LT/MT field (final version)".

4 Data Management Plan – Resource Sharing

The CRACKER Data Management Plan (DMP) provides the data management policy of the project with regard to the produced data sets, containing, among others, information on standards and metadata used as well as on sharing, archiving and preservation. The DMP has been documented in CRACKER Deliverable D3.1, "Data Management Plan", updated in D3.8 and finalized in D3.9. In its final version it also contains the descriptions of all datasets generated by CRACKER or with the support of CRACKER.

The use of a DMP is required for projects participating in the Open Research Data Pilot, which aims to improve and maximise access to and re-use of research data generated by projects. The elaboration of DMPs in Horizon 2020 projects is specified in a set of guidelines applied to any project that collects or produces data. These guidelines explain how projects participating in the Pilot should provide their DMP, i.e., to detail the types of data that will be generated or gathered during the project, and after it is completed, the metadata and standards which will be used, the ways how these data will be exploited and shared for verification or reuse and how they will be preserved.

In principle, projects participating in the Pilot are required to deposit the research data preferably into a research data repository. Projects must then take measures, to the extent possible, to enable for third parties to access, mine, exploit, reproduce and disseminate, free of charge, this research data.

The guidance for DMPs calls for clarifications and analysis regarding the main elements of the data management policy within a project. CRACKER follows the practices of META-SHARE[13] for data documentation, verification and distribution, as well as for curation and preservation, ensuring the availability of the data throughout and beyond the runtime of ICT-17 projects and enabling access, exploitation and dissemination, thereby also complying with the standards of the Open Research Data Pilot.

The respective template identifies, in brief, the following:

Introduction and Scope

Overview and approach on the resource sharing activities underpinning the language technology and machine translation research and development within each participating project and as part of the Cracking the Language Barrier federation.

Dataset Reference and Name

It is recommended that a standard identification mechanism be employed for each data set, e.g., (a) a PID (Persistent Identifier as a long-lasting reference to a dataset) or (b) ISLRN (International Standard Language Resource Number).

---

[13] META-SHARE, www.meta-share.eu and www.meta-share.org, is an open, secure, sharing and exchange network of repositories of language data, tools and related web services, devoted to the sustainable sharing and dissemination of Language Resources (LRs) for the Human Language Technologies domain and other domains where language plays a critical role. META-SHARE was implemented in the framework of the META-NET Network of Excellence, and is designed as a network of distributed repositories of LRs.

Dataset Description

It is recommended that resource and media types are classified as:

- **corpora** (text, audio, video, multimodal/multimedia corpora, n-gram resources),

- **lexical/conceptual resources** (e.g., computational lexicons, ontologies, machine-readable dictionaries, terminological resources, thesauri, multimodal/multimedia lexicons and dictionaries, etc.)

- **language descriptions** (e.g., computational grammars)

- **technologies** (tools/services) that can be used for the processing of data resources

Standards and Metadata

Participating projects have been recommended to deploy the META-SHARE metadata schema for the description of their resources and provide all details regarding their name, identification, format, etc.[14]

Data Sharing

It was recommended that all datasets (including all relevant metadata records) produced by the participating projects would be made available under licenses, which are as open and as standardised as possible, as well as established as best practices. Any interested provider can consult the META-SHARE licensing options and pose related questions to the respective helpdesk.

Archiving and Preservation

As regards the procedures for long-term preservation of the datasets, the participating projects may consider using META-SHARE as an additional, alternative distribution channel for their tools or data sets using one of the following options:

1. As part of the further development and maintenance of the META-SHARE infrastructure, a project that participates in the Cracking the Language Barrier federation may opt to set up its own project or partner-specific META-SHARE repository and link to the META-SHARE network, with CRACKER providing all support necessary in the installation, configuration and set up process.

2. Alternatively, one dedicated Cracking the Language Barrier META-SHARE repository can be set up to host the resources developed by all participating projects, with CRACKER catering for procedures and mechanisms enabling long-term preservation of the datasets.

It should be repeated at this point that following the META-SHARE principles, the curation and preservation of the datasets, together with the rights of their use and possible restrictions, are under the sole control and responsibility of the data

---

[14] Providers of resources wishing to participate in the initiative will be able to request and get assistance through dedicated helpdesks on questions concerning (a) the metadata based LR documentation at helpdesk-metadata@meta-share.eu (b) the use of licences, rights of use, IPR issues, etc. at helpdesk-legal@meta-share.eu and (c) the repository installation and use at helpdesk-technical@meta-share.eu

providers. Alternatively, participating projects may wish to use the META-SHARE repository software for documenting their resources, even if they do not wish to link to the network.

During the project the META-SHARE software was improved in terms of both the underlying software and the data model used by the repository. The Django framework has been upgraded from version 1.4 to 1.7.11, while on the data model level the metadata schema has been revisited in order to allow for more flexibility regarding the description of language resources. Such modifications include relocations and additions of metadata elements and redefinition of controlled vocabularies. The licensing module of the software has undergone a major adaptation including the addition of license values and conditions of use, while a version field is also available to define different versions of the same license. Minor adaptations involved the repository front end, mainly in the search results and resource view sections. Documentation on the new version of the META-SHARE software is provided in CRACKER Deliverable D3.2, "New version of META-SHARE software", and D3.6, "New version of META-SHARE software (update)".

At the time of writing (22 December 2017), 26 datasets that have been produced (test data, training data) by the WMT, IWSLT and QT Marathon events and extended with information on the results of their respective evaluation and benchmarking campaigns (documentation, performance of the systems etc.) are documented and made available through META-SHARE. The corresponding filled in template for each dataset is provided in Appendix 1.

4.1 CEF ELRC

Having taken into consideration the demands of the DoA for strong involvement with the EU translation services as well as a Call for Tenders published by the European Commission (LINK/INFO) to connect the Language Technology R&D&I community to the actual development work of CEF[15] AT and MT@EC[16], CRACKER established communication with the European Language Resource Coordination (ELRC). ELRC has created ELRC-SHARE, a repository powered by META-SHARE dedicated to datasets produced by and for the public administration being, mainly but not exclusively, within the scope of Public Sector Information (PSI). The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF Automated Translation (CEF.AT) platform. The principles for language description and sharing follow the principles also used in the Data Management Plan and adopted for the CRACKER

---

[15]Connecting Europe Facility (CEF) supports the development of high performing, sustainable and efficiently interconnected trans-European networks in the fields of transport, energy and digital services, with the goal of filling in the missing links in Europe's energy, transport and digital backbone. With regard to the digital realm, CEF is meant to facilitate cross-border interaction between public administrations, businesses and citizens.

[16] MT@EC is the European Commission's online machine translation service which produces translations into and from any official EU language. MT@EC is currently available to public administrations in any EU country, Iceland, or Norway, as well as to all EU institutions and agencies and it perform best with texts, or full documents, on EU-related matters.

datasets. The description of the coordination between MT research and CEF has been documented in CRACKER Deliverable D4.5.

5 Joint Events and Dissemination Activities

This section gives a brief description of the joint events and activities in which ICT-17a/b projects participated along with the Cracking the Language Barrier federation or which were held in collaboration with CRACKER. The main goal of participation in a variety of events was to amplify the message disseminated on Multilingual Europe.

5.1 Workshops

The federation, especially its members CRACKER and QT21, organised multiple workshops throughout the project starting from the very successful workshop[17] on "Translation Evaluation: From fragmented tools and data sets to an integrated ecosystem?" at LREC 2016 in Slovenia (on May 28th, 2016).

Another workshop took place on August 6th, 2016, in Berlin with involvement from QT21 and HimL. The main goals were to discuss, with HimL, the work done on Morphology, Semantics and Word Ordering and share the obtained experience. Know-how transfer from one project to the other was also envisaged. The main topics addressed were:

- Domain adaptation paradigms within HimL

- Discussing the QT21 and HimL approaches to morphology

- Discussing the QT21 and HimL approaches to include semantics

- Sharing QT21 experience on word order processing

- Sharing QT21 experience on Automatic Post Editing for domain adaptation

QT21 participated in another workshop, which took place on March 15th, 2017 in Luxembourg with MT@EC. The focus was the specific issues and problems addressed by MT@EC and the transfer of knowledge from QT21 to MT@EC as concerns Neural Machine Translation (NMT), from working principles down to tips and tricks.[18]

The topics addressed at the workshop were:

- Operations and Scalability of NMT

- Training with Nematus on large data volumes

- Decoding optimisation

- Typical NMT issues in real application – translation of numbers and URLs, single word translations, rare word translations

- Domain adaptation with Neural Networks

---

[17] The workshop website including the proceedings is hosted on the website of the federation.
[18] MT@EC has been involved with work on NMT (using Nematus), is in the process of dealing with resources on a much bigger scale than the one used in research and has also dealt with domain adaptation and response time issues.

- Automatic Post Editing (APE) + online APE for domain adaptation

- Dealing with low resources using multi-lingual MT and Pre-trained systems

- Addressing morphology with Byte Pair Encoding (BPE), character based NMT and finite state morphology

HimL, K-Connect, MMT and TraMOOC participated in a workshop, which took place in Valencia on April 4th, 2017 in order to discuss the constraints each project has experienced while going industrial and how important these constraints are in terms of research work. The various topics addressed included among others:

- Listing what users/customers require from MT

- Methods to augment training corpora with automatically generated unseen morphologies

- Differentiating translation for lay-people from translation for experts

- Adapting MT to users by learning iteratively from post edits

- Context analysis for a better targeted terminology

- Use of Direct Assessment for evaluation of domain-specific MT.

5.2   CRACKER and LT_Observatory Common Workshops

The LT_Observatory project has been working towards an online EU-internal market free of language barriers. To this end, it has proceeded towards joining, benchmarking the quality and facilitating the access to language resources – an objective also targeted by CRACKER. This has led to the organization of two workshops.

The first workshop took place in Paris on September 22nd, 2016, and was co-organised with LT-Innovate and ELRA/ELDA attracting stakeholders from various areas within the MT, LR field, such as researchers, companies, public entities, etc. The workshop entitled "Optimising knowledge and practice for operational language resources" provided the ground for discussion on the issues of best practices in LR management (ways to find, create, share, evaluate, use and improve LRs) as well as tools enabling such practices and allowing for rapid expansion of the range of available LRs across more (EU) language pairs. Discussions also revolved around standards for metadata for LRs which in the future might target textual dimensions as semantics, domain-centric, rhetorical, stylistic, etc.

The second workshop was entitled "Collaborating around Language Resources for the Digital Single Market" and it was held in Paris on December 8th, 2016. Again the workshop took place in collaboration with LT-Innovate and ELRA/ELDA. The purpose of the workshop has been to bring together a core group of stakeholders in LRs – i.e., digital parallel resources used for translation identified during the LT_Observatory project to create a plausible agenda of the next three to five years for building intelligent LR management into the MT infrastructure in the EU and use this agenda to explore and define responsibilities, opportunities, collaboration structures, and research – to innovation projects with respect to LR futures. The workshop touched upon key issues which included:

- Building effective relations between the EC's CEF.AT programme and the commercial MT user community for access to existing and new LRs

- Expanding and accelerating LR discovery/creation in Europe

5.3 Presentations

After establishing the federation at META-FORUM 2015, members of the federation participated in META-FORUM 2016 in Lisbon and also in META-FORUM 2017 in Brussels (with poster presentations and/or oral presentations). These events are described in detail in CRACKER Deliverables D1.4, "Report on META-FORUM 2016" and D1.5, "Report on META-FORUM 2017".

Furthermore, the Cracking the Language Barrier federation was presented in regular and invited presentations, some of which include[19]:

- EFNIL 2015 Conference, Helsinki, Finland (8/9 October 2015)

- ICT 2015, Lisbon, Portugal (20/22 October 2015)

- CITIA/META-NET event, Prague, Czech Republic (29/30 October 2015)

- Translating Europe Forum, Brussels, Belgium (27/28 October 2016)

- Future and Emerging Trends in Language Technologies, Machine Learning and Big Data (FETLT), LT_Observatory Final Workshop, Seville, Spain (30 November 2016)

- EC LT Brainstorming Meeting, Luxembourg (13 December 2016)

- EP DG Translation Conference Translation Services in the Digital World, Luxembourg (16/17 October 2017)

- Research Data Workshop, Bonn, Germany (1 December 2017)

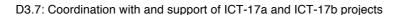- ESCPM Conference, Athens, Greece (1-3 December 2017)

5.4 Shared tasks and evaluations

A series of shared tasks has been created to continuously measure progress and provide a platform for sharing and collaboration. QT21 has been involved in the annual Workshops on Machine Translation (WMT), for which, together with CRACKER, it has been defining the data sets from 2015 to 2018 (WMT16, WMT17 and WMT18). The shared tasks focus on European language pairs with a "guest language" providing a particular translation challenge introduced each year. For 2015-2017 the core language pairs have been English to and from Czech and German. Parallel corpora are provided as training data, and additional resources are available for download. The main goals of WMT have been to:

- investigate the applicability of current MT techniques when translating into languages other than English,

- examine special challenges in translating between European languages, including word order differences and morphology,

---

[19] A detailed presentation of all events is given in CRACKER Deliverable D1.7, "Final Report on Coordination among QT Projects".

- investigate the translation of low-resource, morphologically rich languages,

- create publicly available corpora for machine translation and machine translation evaluation,

- generate up-to-date performance numbers for European languages in order to provide a basis of comparison in future research,

- offer newcomers a smooth start with hands-on experience in state-of-the-art statistical machine translation methods.

All datasets for WMT are described in detail in CRACKER Deliverable D3.9, "Data Management Plan".

CRACKER

Appendix 1

Template for datasets description:

Resource Name	Complete title of the resource
Resource Type	Choose one of the following values: Lexical/conceptual resource, corpus, language description (missing values can be discussed and agreed upon with CRACKER)
Media Type	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
Language(s)	The language(s) of the resource content
License	The licensing terms and conditions under which the LR can be used
Distribution Medium	The medium, i.e., the channel used for delivery or providing access to the resource, e.g., accessible through interface, downloadable, CD/DVD, hard copy etc.
Usage	Foreseen use of the resource for which it has been produced
Size	Size of the resource with regard to a specific size unit measurement in form of a number
Description	A brief description of the main features of the resource (including URL, if any)