MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable 1.3

Final Report

Authors: Bente Maegaard, UCPH, Dorte Haltrup Hansen, UCPH, Khalid Choukri, ELDA

November 2010

PROJECT FINAL REPORT

Grant Agreement number: 214602

Project acronym: MEDAR

Project title: Mediterranean Arabic Language and Speech Technology

Funding Scheme: Support Action, FP7-ICT-2007-1

Date of latest version of Annex I against which the assessment will be made:

24 October 2007.

Period covered: **from** February 1, 2008 **to** July 31, 2010

Name, title and organisation of the scientific representative of the project's coordinator:

Bente Maegaard, professor, University of Copenhagen, Centre for Language Technology

Tel: +45 35 32 90 74

Fax: +45 35 32 90 89

E-mail: bmaegaard@hum.ku.dk

Project website address: http://www.medar.info

1. Final publishable summary report

1.1 Executive summary

MEDAR addresses International Cooperation between Europe and the Arabic Mediterranean region on Speech and Language Technologies. The goals of the MEDAR project are the production and availability of shareable language resources (LRs) and tools for Arabic, the advancement of Arabic language technology, in particular multilingual resources and tools and the cooperation between countries towards these goals. Cooperation is a key concept for the project.

MEDAR builds upon the NEMLAR project, 2003-2005, funded by the European Commission. The MEDAR Consortium consists of 15 partners: 9 from universities, 6 from industry; 7 from the Middle East and 8 from Europe.

MEDAR has 4 main objectives:

1. Developing the **Cooperation Roadmap** based on the foreseeable technological trends, market potentials, and cooperation possibilities.
2. Supporting the **Development of tools and resources** on the basis of partners' technologies and Open Source code for a baseline MT system that will be made available as an MT kit for education, research and Open Source technology development after the project time.
3. Updating the **Basic Language Resource Kit (BLARK) report**: a BLARK is the minimum set of resources and tools necessary for carrying out research and training on LRs and HLT. MEDAR has a focus on MT and MLIR.
4. Consolidating the **Network** of players in all areas of Arabic HLT.

The major outcomes of the project are:

- A cooperation Roadmap that has been widely promoted and hopefully will constitute a corner stone of roadmaps within the Arabic region.
- An example of a concrete implementation of the Cooperation Roadmap, for the field of higher education.
- The 2nd International Conference on Arabic Language Resources and Tools.
- The large number of relations with players outside MEDAR, established through the consortium and the project dissemination activities.
- An MT Baseline package that will be made available.
- An Evaluation package for MT evaluation.
- Evaluation results that establish state of the art.
- A Knowledge Base of players, products & resources and a BLARK updated for Arabic.

Further work should be encouraged and should primarily focus on:

- Taking up other activities as proposed in the Roadmap
- Maintaining the network and relationships established as a discussion forum for new initiatives.
- Updating the Knowledge Base
- Enhancing baseline MT systems

1.2 Summary description of project context and objectives

The development of language resources and tools for the Arabic language is important for the economy in the Arab countries: The industry itself provides new jobs, but more importantly the use of language technology makes administration faster and more efficient. At the same time it is important for the culture. By focussing on Arabic language technology and making both the technology and content available in Arabic, the use of Arabic will grow. Language technology can also help Arab speakers access information in foreign languages, even without a very good knowledge of these languages.

To address this issue MEDAR is structured in three overlapping 'streams': 1) a technical stream, 2) a Cooperation Roadmap stream, and 3) a dissemination stream.

The technical stream covers producing a Knowledge Base on Human Language Technology (HLT, and producing prototypes for machine translation which can be used in universities etc.

BLARK report and Knowledge Base

The project made three surveys that focus essentially on identifying information on players, LRs and tools for Arabic, which then became a part of the BLARK report. The survey identified a large set of requested resources and some available ones. An online Knowledge Base shows the identified players, LR and tools.

A BLARK for any given language usually describes the needs for language resources and tools for the general language and for generic applications. The present BLARK for Arabic is a specific type of BLARK, as it partly describes the needs for general language resources and generic applications, and partly describes the special needs of multilingual language technology, in particular machine translation (MT). Additionally, it does not only give the BLARK specification, but also aims at keeping track of current BLARK content, and such a BLARK report will always be work in progress, and it is hoped that the readers of the BLARK for Arabic will help provide more information so that the document will gradually grow better and more useful. The BLARK for Arabic was published as a report and the BLARK Definition and Content was also implemented online for the sake of accessibility.

As a result of the analysis of the survey findings and the BLARK content, we concluded that even though machine translation for English-Arabic exists on the market, what is needed in order to boost the development, and for initial training of students and young researchers, is Open Source tools and related language resources. Analysis also showed the need for an MT evaluation package.

MT activities

The MT activities covered the selection of MOSES as the building block for an English-Arabic MT and its customization through the use of specific modules (of alignments) and specific data for its training. At the end of the project, the consortium is providing the community with two baselines, a first basic one prepared by the UOB and a more tuned one prepared by IBM together with Dublin City University.

The two baselines have been evaluated within two phases. A first one consisted of (plain) evaluation without supplying any additional resources to enhance the systems capabilities while the second one allowed bringing in more language resources to tune the systems to a new domain. In both phases, test sets have been produced on purpose and have been made available for the participating systems.

In order to conduct such evaluations, a specific evaluation framework has been developed to help run automatic evaluation procedures as well as provide adequate interfaces for human evaluations (subjective assessment by human judges). This platform will be maintained for further evaluations. The collections of resources produced for system training and evaluations have been packaged and will be made available for R&D activities.

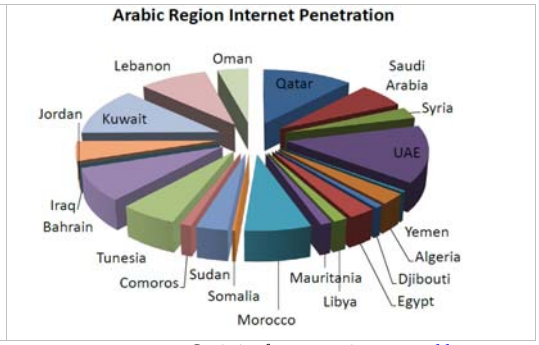
A number of modules have been analysed and the consortium used different ones, for instance different aligners have been used. Some were Open Source and widely used e.g. Champollion, Hunalign (alignment at sentence level) and will be offered as part of the MEDAR package while others, copyrighted by the partners (i.e. CEA) have been extensively used to produce new aligned corpora (at word level) that will be incorporated within the final MEDAR packages.

Last but not least, many partners took this opportunity to work on English-Arabic MT and exploited the baseline provided within the project (e.g. ILSP, RDI), some even did start an MT activity that was not part of their domain of work (e.g. ENSIAS). UCPH was able to finance a side project, focussing on Danish-Arabic, a language pair with very few parallel resources.

Packages made available by the project:

1) MT baseline system(s)
2) Two MT evaluation packages and an MT evaluation platform
3) A set of LRs for MT training
4) A package of corpus aligners

The Cooperation Roadmap stream mainly consists in designing a roadmap for cooperation between players in the EU and Arabic countries, within the Arabic countries, and between academia and industry.



Statistics from www.internetworldstats.com

Cooperation Roadmap

MEDAR has proposed a Cooperation Roadmap with the purpose of building sustainable Human Language Technologies for the Arabic language within and outside the Arabic world.

The roadmap aspires to address a new perspective on collaboration between the Arabic region and the European Union. In order to do so, the MEDAR consortium adopted a multi-dimensional roadmap that combines various impacting factors helping to derive a coherent view. Such factors are related to the state of players, human resources and education curricula, technology development and R&D, evolution of the e-infrastructures (in particular mobile and Internet penetration, attractiveness of ICT environments, growth of amount of e-content in Arabic). Another dimension that is considered is the market: both the domestic and international ones are reviewed and the market profile is analysed (in particular products versus services).

Last but not least, a set of instruments are elaborated upon with the aim to boost cooperation between universities and industries (both within the region but also with the EU and the West in general), to improve the technology transfer (from local R&D players to local/international industries).

The following summarises some components and directions that will lead into success of the strategy:

- Universities and research centres should provide the basic and applied research in cooperation with industry to produce solid products,
- Universities and other educational institutions should create the proper training and re-training (rehabilitation program for personnel from other disciplines who could be re-trained to fit the new requirements)
- Governments and funding agencies should facilitate, support and help companies and universities to initiate and sustain their products,
- Specialized companies should play a significant role in this area and should build and enhance tools, utilities and applications for Arabic HLT,
- Governments should launch services/applications for citizens (e.g. e-government sub-projects, initiatives) that will be accessed and navigated in Arabic language
- International companies specialized and interested in the HLT (and Arabic HLT) in particular should be encouraged:
  - to maintain the interest in Arabic,
  - to provide services to the region and should be given the facilities to make this attractive for them,
  - to maintain relationships with local companies and task forces, and
  - to utilize what is available locally
- Local mobile companies, internet service providers and telephone companies should provide the support and encourage the local companies and universities to direct their efforts towards producing tools and utilities that could be integrated and added to the provided services.

The cooperation roadmap was published as a report and a short version was also produced and translated to Arabic in order to reach a wider audience, both at conferences and when meeting policy makers.

Higher Education cooperation

One of the most important elements for cooperation is the human resources as advances in this field can only be made if a sufficient amount of well trained staff is available. Consequently, university education is a key element, and MEDAR has chosen this as an example in order to see how cooperation can be implemented.

An initial cooperation plan is proposed between four universities (in Lebanon, Egypt, Jordan and Morocco, with University of Copenhagen as a partner) to work jointly on a program that will aim at educating students on all levels (Bachelor, Masters and PhD holders) in the domain of Arabic HLT. The first parts of this plan will be implemented in the academic year 2011-2012. Affiliation and cooperation will also be looked at with similar programs in Europe such as the International Masters in Natural Language Processing and Human Language Technology.
It is foreseen that the cooperation can be extended to many more universities in the same countries and to other countries, once the feasibility and success have been proven.

Network of players

The network of players interested in Arabic HLT has been extended with new players. The various conferences which MEDAR has organised or to which MEDAR partners have been invited have proven extremely useful in establishing new contacts. The closing discussion at the workshop on Semitic languages at LREC2010 gave rise to a group being formed in order to promote some of the roadmap initiatives and other relevant common activities. Consortium partners have been invited to participate in a long range of activities, e.g. meetings in other Mediterranean or UN fora, reviewing for various journals, evaluating research proposals, and academic conferences and seminars. The MEDAR newsletter reaches 230+ readers.

The dissemination stream covers many different types of dissemination as the project had to get in contact with many different target groups: players in all areas of Arabic HLT, decision makers and funding agencies.	

One of the prerequisites for the success of the awareness and dissemination activities is **visibility** and **branding**. MEDAR is following the successful NEMLAR project (2003-2005) and has decided to keep the NEMLAR logo as this is already a known brand. The logo, which combines Arabic and European features, is used for the websites, the flyer, posters, booklets and other publications, PowerPoint presentations etc.

Website, newsletter and information material

The project has acquired the domain www.medar.info for its central website. The website is mainly in English as this is a European project, but important parts, as the project mission, consortium etc., are translated into Arabic. The website contains: *Mission, Consortium, Reports, Conference site, Knowledge Base, BLARK, Publications, Newsletter, Network, Media, Arabic HLT, News, and link to the NEMLAR reports.* The website is continuously updated with relevant news in these categories. The website contains all information relevant to the audiences. However, a website is a 'passive' resource, even if it continuously updated, as it requires the user to go there. Therefore, the project continued the publication of the NEMLAR newsletter which appears every 3 months. The newsletter essentially provides information on the new material which has appeared on the website since last issue. This way, the newsletter serves as the 'active' counterpart of the website, reminding subscribers to go to the website.

Conferences and call for papers was announced on the website and in the newsletter. Similarly, articles in scientific journals and periodicals, international or national, gave an important contribution to the dissemination activities. To reach the general public and policy makers, academic journals and newsletters are, of course, not adequate. To this end, press releases and interviews in daily newspapers and television was used.

In order to be able to provide a short description of the project and contact address etc., the project has produced a flyer containing information on the mission of MEDAR, the working methodology and contact address. The flyer has been distributed at conferences and fairs, given to possible contributors to surveys, to political players and decision makers etc. The flyer is translated into Arabic.

Conferences and workshops

One of the important ways of reaching in particular the academic community was the conference The 2nd International Conference on Arabic Language Resources and Tools organised by MEDAR and taking place April 2009 in Cairo, Egypt. The conference gave an opportunity for external and consortium participants to present their results and to be invited for collaboration on the roadmap.

In connection with the conference in Cairo we arranged 3 tutorials on Arabic NLT and MT with world class teachers. MEDAR also co-organised two workshops on Arabic HLT at the two relevant LREC conferences (2008 and 2010) and. Co-organizing with other parties is an excellent way to start collaboration.

Finally many of the project partners participated in other conferences, workshops and events where a MEDAR presentation was relevant.

1.3 Main S&T results/foregrounds

Surveys and Knowledge Base

An important part of the initial work consisted in surveying the HLT domain for Arabic and ensuring that an up-to-date inventory of players, projects, products, resources, is established and maintained. Three surveys have been conducted as a follow-up of the work carried out within NEMLAR (2003-2005). Several reports were drafted with the description of the surveys' contexts and their results. The major outcome of these is a Knowledge Base which is consolidated as the place to find appropriate information on all issues related to HLT. Such consolidated findings will be accessible through Internet. This Knowledge Base will be maintained and regularly updated regarding the institutions and experts, language technologies and LRs.

The original "survey" and "inventory" works carried out in NEMLAR but also in the first phase of MEDAR, used very extensively the consortium partners' networks to collect the raw data and then went through a compilation and correction stage. The second phase partially used a web-based tool that helped participants to fill in a user-friendly questionnaire, leading to more results. In addition, the consortium exploited more instruments to collect data, in partnership with others. For instance, some new findings were provided by ELRA that authorised the exploitation of new instruments established in conjunction with the EC project FlareNet called LRE-MAP. Such instrument aimed at monitoring the HLT area with regular snapshots, taken at major milestones like LREC Conferences. For instance, the LRE-Map (www.resourcebook.eu) was used to identify resources that were described at LREC2010 and CoLing2010 and that could be unknown to MEDAR.

The consortium benefited also from the work of ELRA LR identification task force that collects data on new resources to identify those related to Arabic. We also used the Semitic workshop (a satellite event of LREC2010), and information collected for the BLARK document and the Cooperation Roadmap. In addition, the consortium did set up a specific website to allow the storage of information of resources usable for MT and MT evaluation.

And finally, the consortium used extensively web-based searching, based on input from new network members, previous results by workshop participants (different from those presented at the workshop) etc. The ACL Specdial Interest Group (SIG) for Semitic Languages with which relationship has been established will also be an important knowledge source for the future.

The list of players, tools, technologies, Language Resources is now part of our Knowledge Base provided as a website, open to the public. Later on additions, updates, corrections will be allowed with moderation taken care of by ELDA and other members of the consortium. This is one of the areas for which collaboration with the ACL SIG can be discussed.

The work carried out within NEMLAR and MEDAR that aimed initially to identify key players within the Arabic region has been extended to the identification of tools, technologies, Language Resources, etc. that are related to Arabic. Over the years, such inventories have been drawn up and exploited to set up partnerships and joint projects. Our goal today is to establish a reliable database that would constitute a trustable reference to all players interested on Arabic HLT. These could be funding agencies, policy makers, research labs, HLT specialized corporations and vendors, researchers, as well as educational institutions.

MT objectives

The objectives of one of the important pillars of MEDAR (the technical pillar, in particular the development of tools and resources for MT) were listed as:

- Developing a framework for the evaluation of English-to-Arabic MT systems;
- Producing data for MT training and evaluation;
- Developing a baseline with background from existing Open Source tools;
- Evaluating MEDAR MT baseline systems and ranking MEDAR baseline MT systems with respect to other MT systems;
- Making available packages containing the full set of resources and tools from MEDAR.
- Creating and federating a new community around the MT English-to-Arabic theme;

The work carried out clearly triggered a useful debate within the HLT for Arabic community. Most of the work carried out so far did focus on Arabic to English and we should admit that this is one of the most tackled pairs. Working on MT for English to Arabic not only highlighted new needs and more balanced view of the HLT community but also more societal discussions about the needs of the Arabic community at large to understand the English writings.

The MEDAR initiative put under strong spotlight the need to establish an appropriate framework for the development of MT technologies and their evaluation as well as the need to enhance the collections of resources required by HLT in general. We feel that the project triggered an area of research that would help new technologies and players emerge within the next coming years.

MT evaluation

This part deals with the evaluation methodology and results of the two MEDAR evaluation campaigns. The context is the evaluation of MT systems for English-to-Arabic direction. The very first goal is to identify the performance level of the baseline systems developed within the project. The evaluation has been conducted in two phases: Phase 1 aiming at setting some basic facts about state of the art for MT on English to Arabic and Phase 2, aiming at collecting enough data to better train and tune the systems and assess the improvements made. A couple of online translation systems have been used to compare with the results submitted by MEDAR participants

As indicated above, the MEDAR evaluation campaigns targeted several objectives, the most important being to develop (and make available) a framework for the evaluation of English-to-Arabic MT systems, to produce the necessary language resources (for training and testing purposes), and (last but not least) to develop an MT baseline system and assess its performance (including in comparison with existing online systems as well as some commercial products available from our partners).

The ultimate goal of this part of MEDAR is to make all this available as package containing the full set of resources and tools from MEDAR.

The first phase of the MEDAR evaluation consisted in developing and packaging two baseline systems (from UOB and IBM/DCU) and then proceeded with the test of the two systems. In order to do so, a test corpus has been built (by UCPH). The test corpus allows scoring the systems against reference translations, produced by human professional translators. The test corpus consisted of English texts selected from a specific domain (Climate Change) and a variety of sources. They are composed of about 10,000 words to be used as a test corpus. The 10,000 words of the test corpus were translated four times by four different translation teams (in Egypt, contact provided by RDI). Specific guidelines were produced, and provided to the translation agencies in order to monitor the translations quality. Specific validation guidelines were also produced for validating these translations.

Translation agencies used their best practices to produce the MEDAR translations. While we trust that each translation agency has its own mechanisms of quality control, we agreed with them on specific "translation guidelines" so that all translations share a common ground. Such guidelines concern the behaviour of the translator who is asked to ensure that target translation is faithful to the original source text in terms of meaning and style (e.g. a press release should be rendered in a journalistic style); the tone and register of the language should be respected (e.g. if the text shows an angry or uneasy speaker in the source language, this state of mind should be also expressed in the target language, conveying the same tone), etc. We also had our own validation guidelines to assess (a posteriori) the quality of the translations according to the translation guidelines. The validation guidelines allow the validator to score the errors discovered within the translation and then allow us to draw conclusion about accepting or rejecting the translations (if rejected a translation had to undergo a new translation phase).

The validation consisted of automatic validation (e.g. checking the formats) and human validations that aimed at identifying syntactic, lexical, and other types of errors.

During the first phase (first evaluation), the two baseline SMT systems were tested. As indicated above, these had been developed by the University of Balamand and IBM/DCU on the basis of MOSES.

In addition, the evaluation campaign was open to external participants (non members of the MEDAR consortium). Many of the key players in English-Arabic MT (but more generally SMT) expressed their interest. Due to the short period of time to execute such evaluation and also due to the fact that many players did not assess the effort needed to develop English to Arabic component (often on the basis of an existing Arabic-English MT system), most of them declined our offer to participate at the last minute. For comparison purposes, two online systems have been used in this evaluation: Google Translate and Systranet. Their results must be considered carefully since they are not really participating systems and the R&D groups at these two players may have better systems for this pair than the ones online.

There was no training or development phase planned for the first MEDAR evaluation campaign, therefore no data was provided to participants. The two MEDAR baseline systems had not been specifically trained, a very basic data set had been used (this is a small corpus included in each package).

In order to evaluate the MT systems performances, we re-used some of the well known scoring tools (used within TC-STAR, CESTA and known in the framework of MT DARPA/NIST evaluation), in particular BLEU, BLEU/NIST and mWER metrics. (BLEU, which stands for BiLingual Evaluation Understudy, counts the number of word sequences (n-grams) in a sentence to be evaluated, which are common with one or more reference translations. BLEU/NIST, is a variant metric of BLEU, which applies different weight for the n-grams, functions of information gain and length penalty and mWER, Multi reference Word Error Rate, computes the percentage of words which are to be inserted, deleted or substituted in the translated sentence in order to obtain the reference sentence.

The Results have been automatically computed against the four references. To compare to what a human translator can produce, we did consider one of the translators (chosen arbitrarily) as an "MT system" (in principle a perfect one!!). As a highlight, the human translator achieves about 72.5%of correct words (27.5 error rate) while the first system achieves less then 35*% of correct words (66% of errors).*

These automatic measures show very low performances at that point. This campaign may be considered as a dry run, so as to test the protocol and the organization and establish the baseline instead of testing the systems objectively. The low scores should also be interpreted carefully, it is clear that the test corpus is difficult for translation, even for a professional translator. On this basis, the results are not as bad as they may look.

Finally, we can argue that BLEU or any current automatic metrics may be not adapted to process Arabic data (n-grams not suitable to an agglutinative language).

As planned, a second evaluation campaign was initiated and aimed at testing the systems after tuning. Therefore, training data was provided to increase the size of the training corpus.
Two training conditions were considered: Constrained Training (only data provided by MEDAR could be used to train the systems) and Unconstrained Training (no restriction with respect to the data that may be used to train the MT systems).
Various language resources have been compiled either as monolingual corpus (to train the monolingual language modelling component) or bilingual and aligned to train the alignment component. All these resources will be made available to the community as the corpus package from MEDAR.
For testing purposes, an evaluation corpus (English texts) was selected from the same domain (Climate Change). Again about 10,000 words were used as a test corpus. The preparation of the evaluation data has been done in the same way as for the first MEDAR evaluation campaign.

Several submissions were allowed per participant, one of them had to be labelled as the "primary" submission. Others are considered as "secondary" submissions.

As for the first MEDAR evaluation campaign, two online systems have been used in this evaluation: Google Translate and Systranet. A submission was made by the University of Columbia (external participant). Other systems were made by ENSIAS, UOB, IBM/DCU and Sakhr. A total of six submissions have been made. All the submissions have been done within the Constrained Condition which clearly reflects the gaps in resources to train such tools outside the ones provided by the project.

Computation of the scores used the same techniques and tools exploited within the first MEDAR evaluation.

The results of the second MEDAR evaluation campaign did not show any improvement, including for the two online systems. These results are quite surprising since more training data were provided to the systems. The results of one of the human reference translations (considered as an MT system) are slightly higher than for the first evaluation, which would mean that the test corpus was more easy to translate (or more likely that our translator of phase II is more skilled, which we can not check).
Lower scores of the baseline systems can be explained by the large number of non recognized words (words out of the system vocabulary and absent from its training data). Most of the low quality translations consist of transliterated word (words unknown to the system).

The performance within MEDAR is still too low compared to current systems using similar approaches for other language pairs. A number of open issues have to be tackled in order to improve the performance:

1. Increase the size of training data.

2. Incorporate more tools to account for the specific features of Arabic. We have noticed that preprocessing proved to be efficient.

3. Ensure that the scoring metrics are appropriate for assessing Arabic outputs (e.g. BLEU measures some "consistencies" of n-grams, it may not be adapted to an agglutinative language like Arabic). Human evaluation will be conducted to check this issue. Results of such an evaluation (for instance using Fluency or Adequacy criteria) would allow us to compare human and automatic metrics within a "meta-evaluation" (i.e. the evaluation of the metrics).

4. Improve MOSES for Arabic, adding specific features e.g. reordering words for alignment, syntactic analysis for preprocessing, segmentation and morphological decomposition, word alignment, etc.

Despite the low performance achieved by several systems based or derived from MOSES, MEDAR will offer these packages to the HLT community. These contain the two baseline systems and the corresponding training and evaluation resources.

Support for the Roadmaps

In addition to the achievements mentioned above (related to MT activities), a main outcome of the project is the MT packages that will be offered to the HLT community having an impact on research and training carried out at several universities in the Arabic countries. The plans and schedules are part of the education roadmap that has been designed within the project.

1.4 The potential impact and the main dissemination activities

Impact

One evident and significant success case has been realized by February 2010 regarding the adoption of a roadmap for HLT in Egypt where the semi governmental IT Industrial Development Agency (ITIDA) has approved the initial funding of the new NGO foundation Arabic Language TEchnology Center (ALTEC) that adopts an Egyptian version of the MEDAR roadmap.

This is exactly what the project aims at: the adoption of the MEDAR roadmap by individual countries, or the adoption of a modified version hereof. The Roadmap is made for the whole region, and one cannot expect it to fit all countries equally well. So, one potential impact of the project is that policy makers and university staff adopt the ideas in the roadmap report, by starting to implement some of them, by finding partners through the network who have similar ideas etc. Egypt has shown the way, and we believe that more players and countries will follow. Discussions held at various conferences and other meetings have been very positive, and the project believes that we have given the local players and the community as such a very useful tool for guiding the future.

A very concrete impact of the project is the planned higher education collaboration between four Arabic universities from the consortium (in Lebanon, Egypt, Jordan and Morocco, with University of Copenhagen as a partner) to work jointly on a program that will aim at educating students on all levels (Bachelor, Masters and PhD students) in the domain of Arabic HLT. The first part of the plan will be implemented in the academic year 2011-2012.

Another way to measure the impact of the project is to see how many is quoting the project or project activities. According to *bizinformation.dk* 1365 domains are linking to the project website: www.medar.info. And when doing a Google search for *medar language technology* 78 out of the first 100 hits are about the MEDAR project or its activities (excluding the website internal links), which indicates that the MEDAR brand is wide spread.

The consortium has achieved one of its major goals that consisted in customising MT prototypes based on state of the art, Open Source software, MT package MOSES. This package will be made widely available with its accompanying documentation, in particular to the academic players within the Arabic world (and beyond). We expect that such dissemination will boost R&D activities and set up the ground for activities on MT and NLP within a large number of centres. We consider that providing an MT kit with all necessary items to quickly run a baseline would help many academic centres to offer courses and training at all levels (Master, PhD) at very low cost.

In addition, the MT-evaluation campaigns allowed to better understand the low level of current performances on English-Arabic MT systems. The best results are still beyond 2/3 of error rates (with a metric based on mWord Error Rate, mWER=66% for the best systems). This also points out the need to deeply investigate the behaviour of assessment measures when coping with Arabic output. The adaptation of BLEU (and other measures based on n-grams similarities) requires more research that some of the academic partners will be conducting within their own R&D agenda.

The experiments with Danish-Arabic MT which took place at UCPH, based on other funding, will have two types of impact: first of all it may contribute to the development of a methodology for language pairs with very few parallel resources (there is a good contact to the Columbia University on this), and secondly it will have the same impact as the English-Arabic mainstream MEDAR: it can be used for education and training, and some day it may provide good enough results.

Main dissemination activities

The most significant dissemination of the project was The 2nd International Conference on Arabic Languages and Tools, held 22- 23 April 2009 in Cairo, Egypt. It was a follow-up of the 1st International Conference on Arabic Languages and Tools arranged by the NEMLAR project consortium in September 2004.

The conference received a high number of submissions, and the conference welcomed more than 130 participants, incl. the press. The participants came from 25 different countries in four continents, with 60 participants from Egypt and 21 from other Arabic speaking countries in the region. The conference fee was set very low for students, because students "build the future", and the conference participant distribution shows that many students took advantage of this: 39% students, 41% academic participants and 20% industrial participants. We think this is a very nice mix with also a considerable industrial participation.

Cairo University and MEDAR in collaboration provided three free tutorials, with more than 90 students and some faculty members, which gave a world class introduction to Arabic NLP. The three tutorials were:

- Nizar Habash, Columbia University: *Introduction to Arabic Natural Language Processing*
- Mona Diab, Nizar Habash, Columbia University: *Arabic Dialect Processing*
- Andy Way, Dublin City University, Hany Hassan, IBM Cairo: *Statistical Machine Translation: Trends & Challenges*

The conference proved to be a very good forum for discussion and planning of further collaboration. The remarkable success of both the first and the second international conference on Arabic LRs, tools and evaluation, regarding the quality of participations, the attendees, the organization, the sponsorships, the press coverage, and the clear wish in the region, are all encouraging the stakeholders to pursue the potential of holding this unique event devoted to Arabic HLT on a regular basis.

The program and proceedings can be found at the conference website.

MEDAR also co-organised 2 workshops on Arabic HLT in the period:

- HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects, Marrakesh, Morocco 2008
- Workshop in LR and HLT for Semitic Languages, Valletta, Malta 2010.

Beside these big events MEDAR partners participated in various conferences and workshops, were invited to give talk on MEDAR, review and edit papers and journals and published papers. MEDAR was also present with a stand in ICT 2008 International Village and at LREC 2010 EC Village which gave the project the possibility to talk to a large amount of interested people. For these events MEDAR made a flyer (in English and in Arabic) and the Cooperation Roadmap in a short version (in English and in Arabic).

MEDAR has been present in the mass media on several occasions. MEDAR partners were interviewed live on the project via the most impactful Arabic satellite channel Al-Jazeera www.Aljazeera.net during a morning prime time on Aug. 19th, 2009, as well as the Egyptian TV channels: Nile TV and Egyptian TV Channel 1, and Jordan TV Morning Daily Show. The Al-Jazeera interview can be seen

or downloaded from the MEDAR website. The project and its activities have also frequently been covered by many other popular as well as IT-specialized web portals, magazines, and newspapers published in the Arab region; e.g. Lughat Al-Assr magazine, and Moheet portal.

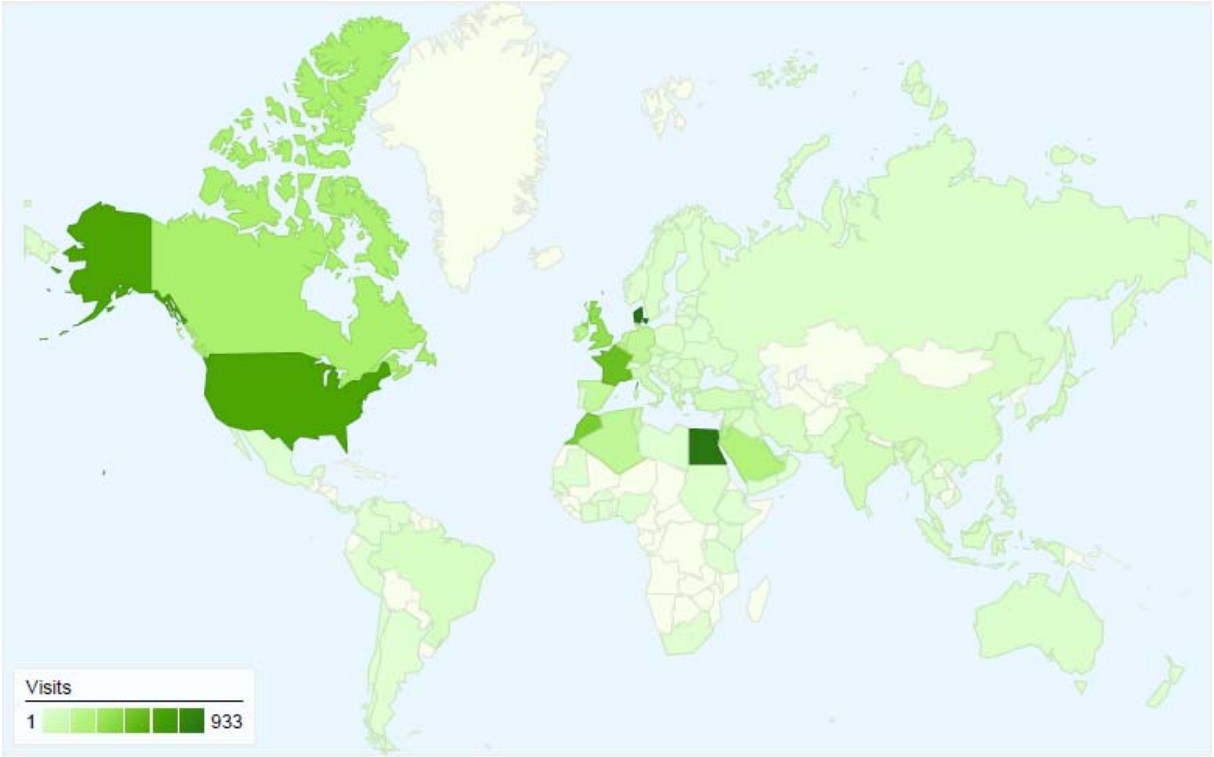
For continuous dissemination the project website (www.medar.info) contains information on news and events in the area of Arabic HLT, an online BLARK for Arabic, a Knowledge Base on Arabic HLT, a quarterly newsletter, a list of scientific publications by the project partners, project activities and reports. The website has been well visited by visitors from 110 countries.

Summary of Analytics for the medar.info website:

Visits in the period (1/5 2009-31/7 2010)	7427
Avg. visits pr. month	495
Avg. visits pr. month (range)	336 - 948
Countries	110
Visits from non-partner countries	4125
Visits from Arabic speaking countries	1869
New visits	64,05 %
Avg. time on the site	00:02:52
Avg. page/visit	2,63
Most busy periods	8/5 -19/5 2009, 4/5 2010
Avg. Bounce rate	59,05%

Google ranking (1/8 2010):
Search term: *Arabic Language Technology* (No. 1 hit out of **8.160.000** hits)
Search term: *Arabic HLT* (No. 2 hit out of **66.400** hits)
Search term: *medar* (No. 3 hit out of **1.940.000** hits)
Search term: *Arabic NLP* (No. 4 hit out of **1.440.000** hits)

The visits came from 110 countries:



The NEMLAR Newsletter is the active counterpart of the website. The purpose of the newsletter is partly to tell about the activities of the MEDAR project, partly to announce events and activities in the field of Arabic language technology. 230 persons subscribe to the newsletter.

Finally, MEDAR established a network group in LinkedIn for people to be able to share ideas, tools and resources and be able to get in contact with each other. The ACL SIG (Special Interest Group) on Semitic languages, however, is very interested in joining forces with MEDAR, and this is a proposal which is being considered carefully.

The project website:
www.medar.info.

The conference website:
2nd International Conference on Arabic Language Resources and Tools

The LinkedIn network group:
NEMLAR Network

The MEDAR Consortium:

Coordinator:
University of Copenhagen, Denmark
Bente Maegaard, Dorte Haltrup Hansen, nemlar@hum.ku.dk

ELDA - Evaluations and Language resources Distribution Agency, France
Khalid Choukri, choukri@elda.org

University of Balamand, Lebanon
Chafic Mokbel, chafic.mokbel@balamand.edu.lb

Al-Ahlyya Amman University, Jordan
Mustafa Yaseen, mustafa@ats-ware.com

Universiteit Utrecht, The Netherlands
Steven Krauwer, s.krauwer@uu.nl

ILSP – "Athena" Research Centre, Greece
Stelios Piperidis, spip@ilsp.gr

RDI - The Engineering Company for Development of Computer Systems development , Egypt
Mohamed Attia, m_atteya@rdi-eg.com

Birzeit University, West Bank and Gaza Strip
Kanan Al-Ali, Kananalali@yahoo.com

ENSIAS - University of Mohammed V Soussi, Morocco
Abdelhak Mouradi, Mouradi@ensias.ma

CEA - Commissariat à l'Energie Atomique, France
Nasredine Semmar, Nasredine.semmar@cea.fr

CNRS - Centre Nationale de la Recherche Scientifique, France
Fathi Debili, fathi.debili@wanadoo.fr

The Open University, United Kingdom
Anne Deroeck, A.Deroeck@open.ac.uk

Université Lumière Lyon 2, France
Joseph Dichy, Joseph.Dichy@univ-lyon2.fr

IBM - Human Language Technologies Group, Egypt
Ossama Emam, emam@eg.ibm.com

Sakhr Software Company, Egypt
Hamdy Soliman, hamdys@sakhr.com

2 Use and dissemination of foreground

Section A (public)

LIST OF SCIENTIFIC (PEER REVIEWED) PUBLICATIONS, STARTING WITH THE MOST IMPORTANT ONES

NO.	Title	Main author	Title of the periodical or the series	Number, date or frequency	Publisher	Place of publication	Year of publication	Relevant pages	Is/Will open access[1] provided to this publication?
1	Proceedings of the Second International Conference on Arabic Language Resources and Tools	B.Maegaard (ed.)		April, 2009	The 2nd International Conference on Arabic Language Resources and Tools http://www.MEDAR.info/Conference_All/2009/index.php	Cairo, Egypt	2009	124 pages	yes
2	Proceedings of HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects	K. Choukri (ed.)		May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008	121 pages	yes
3	Proceedings of the workshop in LR and HLT for Semitic Languages	K. Choukri (ed.)		May, 2010	7th Int'l. Conference on Language Resources and Evaluation, LREC2010 http://www.lrec-conf.org/lrec2010	Paris, France	2010	124 pages	yes
4	Cooperation for Arabic Language Resources and Tools – The MEDAR Project	B.Maegaard	Proceedings of the workshop in LR and HLT for Semitic Languages	May, 2010	7th Int'l. Conference on Language Resources and Evaluation, LREC2010 http://www.lrec-conf.org/lrec2010	Paris, France	2010		yes
5	Towards Networking Efforts to Build an Arabic Cooperation Roadmap	M. Yaseen	Proceedings from The first International Conference on Networked	July, 2009	Published and indexed by IEEE Xplore	Ostrava, The Czech Republic	2009	p. 441-447	no

[1] Open Access is defined as free of charge access for anyone via Internet. Please answer "yes" if the open access to the publication is already established and also if the embargo period for open access is not yet over but you intend to establish open access afterwards.

			Digital Technologies (NDT 2009),						
6	MEDAR: Creating a framework for Arabic language technology	M. Attia	Multilingual Computing & Technology magazine	July/Aug. 2010	Multilingual Computing & Technology magazine www.Multilingual.com	USA	2010		yes
7	MEDAR: Arabic Language Technology, State-of-the art and a Cooperation Roadmap	B.Maegaard	Proceedings of the Second International Conference on Arabic Language Resources and Tools	April, 2009	The 2nd International Conference on Arabic Language Resources and Tools http://www.MEDAR.info/Conference_All/2009/index.php	Cairo, Egypt	2009		yes
8	MEDAR – collaboration between European and Mediterranean Arabic partners to support the development of language technology for Arabic	B.Maegaard	Proceedings of the 6th International Conference on Language Resources and Evaluation	May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008		yes
9	Accessibility and Enablement Through A MEDAR-Cooperation Roadmap for Arabic Language	M. Yaseen	Internet Governance Forum;*Equality in access to knowledge society through language and cultural diversity*	November 2009	IDRC (of Canada); ACSIS and Bibliotheca Alexandrina	Sharm El Sheikh, Egypt	2009		no
10	Using English as a Pivot Language to Enhance Danish-Arabic Statistical Machine Translation	M. Al-Hunaity	Proceedings of the workshop in LR and HLT for Semitic Languages	May, 2010	7th Int'l. Conference on Language Resources and Evaluation, LREC2010 http://www.lrec-conf.org/lrec2010	Paris, France	2010		yes
11	Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons	N. Semmar	Proceedings of the Semitic Languages Workshop - LREC 2010 Conference	May 2010	7th Int'l. Conference on Language Resources and Evaluation, LREC2010 http://www.lrec-conf.org/lrec2010	Paris, France	2010		yes
12	Autonomously Normalized Horizontal Differentials as Features for HMM-Based Omni Font-Written OCR Systems for Cursively Scripted Languages	M. Attia	Proceedings from IEEE International Conference on Signal & Image	November 2009	IEEE International Conference on Signal & Image Processing Applications	ICSIPA09 Kuala Lumpur-Malaysia	2009	p. 185 - 190	yes

			Processing Applications						
13	Arabic Language Resources and Tools for Speech and Natural Language: KACST and Balamand	C. Mokbel	Proc. of the 2nd International Conference on the Arabic Language Resources and Tools	April 2009	The 2nd International Conference on Arabic Language Resources and Tools http://www.MEDAR.info/ Conference_All/2009/index.php	Cairo, Egypt	2009		yes
14	Broadcast News Transcription Baseline System using the NEMLAR database	C. Mokbel	Proceedings of the HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects	May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008		yes
15	A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields	M. Attia	Proceedings of the 6th International Conference on Language Resources and Evaluation	May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008		yes
16	Can the building of corpus-based Arabic concordances with AraConc and DIINAR.1 tackle the issue of Arabic polyglossia?	J. Dichy	Proceedings of the HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects	May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008		yes
17	Automatic versus interactive analysis for the massive vowelization, tagging and lemmatization of Arabic	F. Debili	Proceedings of the HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects	May, 2008	6th Int'l. Conference on Language Resources and Evaluation, LREC2008 http://www.lrec-conf.org/lrec2008	Paris, France	2008		yes

LIST OF DISSEMINATION ACTIVITIES								
NO.	**Type of activities[2]**	**Main leader**	**Title**	**Date**	**Place**	**Type of audience[3]**	**Size of audience**	**Countries addressed**
1	Conference	B. Maegaard, K. Choukri	2nd International Conference on Arabic Language Resources and Tools	22-23 April, 2009	Cairo, Egypt	Research, Industry, Students	130	25 countries
2	Interview	M. Yaseen	Interview on AlJazeera Satellite Channel	19 August, 2009				Arabic Region
3	Workshop	B. Maegaard, K. Choukri	LR and HLT for Semitic Languages	17 May, 2010	Valetta, Malta	Research, Students	53 participants	18 different countries
4	Workshop	B. Maegaard, K. Choukri	HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects	31 May, 2008	Marrakech, Morocco	Research, Students	43	15 countries
5	Tutorials	B. Maegaard	T1: Introduction to Arabic NLP T2: Arabic Dialect Processing T3: MT: challenges and trends	21 April, 2009	Cairo, Egypt	Students	90	Egypt and Arabic region
6	Exhibition stand	B. Maegaard, D.H. Hansen	EC Village at LREC 2010	19 -21 May 2010	Valetta, Malta	Research, Industry, Students		
7	Exhibition stand	B. Maegaard, D.H. Hansen	International Village at the ICT Conference	25 – 27 November 2008	Lyon, France	Research, Industry, Policy makers		
8	Presentation	B. Maegaard	*Digital Libraries and Content – Challenge*	5-7 December 2009	International Networking			

[2] A drop down list allows choosing the dissemination activity: publications, conferences, workshops, web, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters, Other.

[3] A drop down list allows choosing the type of public: Scientific Community (higher education, Research), Industry, Civil Society, Policy makers, Medias ('multiple choices' is possible.

			objectives		Event on Med-EU Information and Communication Technology Cooperation, Royal Scientific Society, Amman, Jordan			
9	Presentation	B. Maegaard	*The MEDAR Project (Objectives, Achievements, Opportunities)*	18-19 June 2009	INCO-NET MIRA Istanbul, Turkey			
10	Presentation	A. Mouradi	*Outils et Ressources linguistique pour la langue Arabe dans le cadre des Projets européen NEMLAR et MEDAR*	9 June, 2009	Rabat, Morocco			
11	Presentation	J. Dichy	The MEDAR Euro-Mediterranean project, a short overview	26-28th April 2009	Damascus, Syria			
12	Presentation	B. Maegaard	MEDAR – a Euro-Mediterranean collaboration for Arabic	25 – 27 November 2008	Lyon, France	Research, Industry, Policy makers		
13	Presentation	C. Mokbel	Arabic Speech and Language processing: Tools and Resources	April, 2008	Expert group meeting on Promoting Digital Arabic Content in the ESCWA Region			
14	Interview	Mohsen Rashwan		22 April 2009	Egyptian TV: Channel 1	General Public		
15	Interview	M. Yaseen		April, 2009	Jordan TV Morning Daily Show			Jordan
16	Interview	M. Attia		July 2009	Al-Arab Al-Qatariyya (daily news paper)	General public		
17	Interview	Mohamed Attia		23 April 2009	Lughat Al-Assr magazine	People with IT interests/background		
18	Interview	M. Rashwan		October 2009	Moheet (Arabic web			

					portal based in Egypt)			
19	Interview	The MEDAR Consortium	The Breakthrough of Arabic Language Technologies	24 April, 2009	Islam Online (IOL) In English			Arabic Region
20	Interview	Ahmed Ragheb		May, 2008	NILE TV (an Egyptian TV channel)	General Public		
21	Interview	B. Maegaard	"At digigalisere arabisk"	April, 2009	Humanist, Copenhagen, Denmark	Research, Students		Denmark
22	Poster	B. Maegaard		5-7 December 2009	International Networking Event on Med-EU Information and Communication Technology Cooperation, Royal Scientific Society, Amman, Jordan			
23	Website	D.H. Hansen	www.medar.info	Ongoing	www.medar.info	Research, Industry, students Policy makers	App. 500 visits a month	110 countries visited the site
24	Flyer	B. Maegaard	MEDAR Mediterranean Arabic Language and Speech Technology		www.medar.info	Research, Industry, Students, general public		
25	Booklet	B. Maegaard	Cooperation roadmap – short version	May, 2010	www.medar.info	Research, Industry, Policy makers		
26	Handout	B. Maegaard	Summary of the General discussion of the Workshop on LR and HLT for Semitic Languages,	July, 2010	www.medar.info	Research, Industry, Policy makers		

Section B -Part B1

TEMPLATE B1: LIST OF APPLICATIONS FOR PATENTS, TRADEMARKS, REGISTERED DESIGNS, ETC.					
Type of IP Rights:	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Application reference(s)	Subject or title of application	Applicant (s) (as on the application)
	NONE				

Section B - Part B2

Type of Exploitable Foreground[4]	**Description of exploitable foreground**	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	**Exploitable product(s) or measure(s)**	**Sector(s) of application**[5]	**Timetable, commercial or any other use**	**Patents or other IPR exploitation (licences)**	**Owner & Other Beneficiary(s) involved**
MEDAR MT kits	MT baselines	NO	None	MT Kits	HLT	for non for profit use (R&D,Commercial organizations engaged in R&D)	Licensable via ELRA	Owner = MEDAR
MEDAR MT Evaluation packages	Evaluation packages	NO	None	Evaluation packages	HLT	for non for profit use (R&D,Commercial organizations engaged in R&D)	Licensable via ELRA	Owner = MEDAR
MEDAR LRs	Language Resources collections for MT & NLP	NO	None	Language Resources	HLT	for non for profit use (R&D,Commercial organizations engaged in R&D)	Licensable via ELRA	Owner = MEDAR

[19] A drop down list allows choosing the type of foreground: General advancement of knowledge, Commercial exploitation of R&D results, Exploitation of R&D results via standards, exploitation of results through EU policies, exploitation of results through (social) innovation.

[5] A drop down list allows choosing the type sector (NACE nomenclature) : http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

In addition to the table, please provide a text to explain the exploitable foreground, in particular:

The outcomes of the project consist of these three kits:

- MEDAR MT kits (MT baselines)
- MEDAR MT Evaluation packages
- MEDAR LRs Language Resources collections for MT & NLP

- *Its purpose*

The consortium, capitalizing on its experience within MEDAR and NEMLAR, will ensure that such packages are made available to the HLT Community. It is expected that the MT kits will trigger new initiatives, in particular within the academic institutions of the Arabic world that will have access to that "start" kit. The MT evaluation packages will be useful to a larger community, for instance the players that could not join the official evaluations. The most important result of the project (in terms of foreground) is the Language Resources produced, compiled, and packaged. These will be made widely available through the usual channels of ELRA.

- *How the foreground might be exploited, when and by whom*

The foreground is part of the NEMLAR/MEDAR assets and will be exploited jointly by the consortium. The consortium has extensive expertise in doing so through the assets developed within the NEMLAR project. ELRA acted as the distribution centre and distributed over 17 copies of NEMLAR resources (13 copies of the NEMLAR Corpus, 3 copies of the NEMLAR broadcast news data and a copy of the speech synthesis database). The revenues generated by the distribution of such resources are managed by UCPH and were used to cover expenses of students attending major conferences and workshops on Arabic & HLT. The developed foreground has been packaged during the life of the project and is ready for exploitation.

- *IPR exploitable measures taken or intended*

As indicated above, all IPR issues have been cleared during the life of the project and the MEDAR Kits will be made available through clean and clear licensing schema (inspired from ELRA & Creative Commons). The licenses related to the MT baselines will be those defined by the MOSES licences (GNU LGPL Version 3, 29 June 2007). Such licences allow us to redistribute copies of the MEDAR MT Kits. For Language Resources and the evaluation packages, specific agreements have been negotiated as part of ELRA's usual activities and will allow distributing the final version.

- *Further research necessary, if any*

Not applicable

- *Potential/expected impact (quantify where possible)*

As indicated above, we do expect a number of potential players to adopt the MEDAR MT Kit (based on MOSES) for their own activities. We expect more than a dozen of players to acquire both the LR package and the evaluation packages over the next 3 to 5 years.

This availability will have an impact on several research areas and particular boost the evaluation initiatives going on through the integration of English to Arabic language pair. The availability of Language Resources will also impact the work on MT in general.

3 Report on societal implications

A General Information *(completed automatically when **Grant Agreement number** is entered.*

Grant Agreement Number:	214602
Title of Project:	Mediterranean Arabic Language and Speech Technology
Name and Title of Coordinator:	Professor Bente Maegaard

B Ethics

1. Did your project undergo an Ethics Review (and/or Screening)? • If Yes: have you described the progress of compliance with the relevant Ethics Review/Screening Requirements in the frame of the periodic/final project reports? Special Reminder: the progress of compliance with the Ethics Review/Screening Requirements should be described in the Period/Final Project Reports under the Section 3.2.2 *'Work Progress and Achievements'*	***No***
2. Please indicate whether your project involved any of the following issues (tick box) :	***No***
RESEARCH ON HUMANS	
• Did the project involve children?	
• Did the project involve patients?	
• Did the project involve persons not able to give consent?	
• Did the project involve adult healthy volunteers?	
• Did the project involve Human genetic material?	
• Did the project involve Human biological samples?	
• Did the project involve Human data collection?	
RESEARCH ON HUMAN EMBRYO/FOETUS	
• Did the project involve Human Embryos?	
• Did the project involve Human Foetal Tissue / Cells?	
• Did the project involve Human Embryonic Stem Cells (hESCs)?	
• Did the project on human Embryonic Stem Cells involve cells in culture?	
• Did the project on human Embryonic Stem Cells involve the derivation of cells from Embryos?	
PRIVACY	
• Did the project involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	
• Did the project involve tracking the location or observation of people?	
RESEARCH ON ANIMALS	
• Did the project involve research on animals?	
• Were those animals transgenic small laboratory animals?	
• Were those animals transgenic farm animals?	
• Were those animals cloned farm animals?	
• Were those animals non-human primates?	
RESEARCH INVOLVING DEVELOPING COUNTRIES	
• Did the project involve the use of local resources (genetic, animal, plant etc)?	
• Was the project of benefit to local community (capacity building, access to healthcare, education etc)?	
DUAL USE	
• Research having direct military use	
• Research having the potential for terrorist abuse	

C Workforce Statistics

3. Workforce statistics for the project: Please indicate in the table below the number of people who worked on the project (on a headcount basis).

Type of Position	Number of Women	Number of Men
Scientific Coordinator	1	
Work package leaders	1	2
Experienced researchers (i.e. PhD holders)	2	20
PhD Students	1	4
Other	8	19

4. How many additional researchers (in companies and universities) were recruited specifically for this project?	**5**
Of which, indicate the number of men:	3

D Gender Aspects

5. Did you carry out specific Gender Equality Actions under the project?	○ X	Yes No

6. Which of the following actions did you carry out and how effective were they?

		Not at all effective				Very effective
❑	Design and implement an equal opportunity policy	○	○	○	○	○
❑	Set targets to achieve a gender balance in the workforce	○	○	○	○	○
❑	Organise conferences and workshops on gender	○	○	○	○	○
❑	Actions to improve work-life balance	○	○	○	○	○
X	Other:	None				

7. Was there a gender dimension associated with the research content – i.e. wherever people were the focus of the research as, for example, consumers, users, patients or in trials, was the issue of gender considered and addressed?

○ Yes- please specify

X No

E Synergies with Science Education

8. Did your project involve working with students and/or school pupils (e.g. open days, participation in science festivals and events, prizes/competitions or joint projects)?

X Yes- please specify
We held 3 tutorials for student at Cairo University and gave 6 travel grants to student to come to our conference in Cairo

○ No

9. Did the project generate any science education material (e.g. kits, websites, explanatory booklets, DVDs)?

X Yes- please specify
We have made an MT tool kit (en-ar) and a BLARK report which are directed towards education and research

○ No

F Interdisciplinarity

10. Which disciplines (see list below) are involved in your project?

1.1 Software development

5.4 Linguistics

X Main discipline[6]:

○ Associated discipline[6]: ○ Associated discipline[6]:

G Engaging with Civil society and policy makers

11a Did your project engage with societal actors beyond the research community? *(if 'No', go to Question 14)* Yes, we had discussions with Presentation at UN-ESCWA.The aim was: know what research to be done, disseminate results	X	Yes No

[6] Insert number from list below (Frascati Manual).

11b If yes, did you engage with citizens (citizens' panels / juries) or organised civil society (NGOs, patients' groups etc.)?

X No
○ Yes- in determining what research should be performed
○ Yes - in implementing the research
○ Yes, in communicating /disseminating / using the results of the project

11c In doing so, did your project involve actors whose role is mainly to organise the dialogue with citizens and organised civil society (e.g. professional mediator; communication company, science museums)?	○ X	Yes No

12. Did you engage with government / public bodies or policy makers (including international organisations)

- MEDAR presented and discussed in different meetings with UN-ESCWA.
- Contact with ITIDA (Egyptian IT Industry Development Agency) www.ITIDA.gov.eg for both the cooperation roadmap and for the organization of the MEDAR conference.

○ No
○ Yes- in framing the research agenda
○ Yes - in implementing the research agenda
○ Yes, in communicating /disseminating / using the results of the project

13a Will the project generate outputs (expertise or scientific advice) which could be used by policy makers?

X Yes – as a **primary** objective (please indicate areas below- multiple answers possible)
○ Yes – as a **secondary** objective (please indicate areas below - multiple answer possible)
○ No

13b If Yes, in which fields? Information Society**,** Education, Training, Youth

Agriculture		Energy		Human rights	
Audiovisual and Media		Enlargement		Information Society	
Budget		Enterprise		Institutional affairs	
Competition		Environment		Internal Market	
Consumers		External Relations		Justice, freedom and security	
Culture		External Trade		Public Health	
Customs		Fisheries and Maritime Affairs		Regional Policy	
Development Economic and Monetary Affairs		Food Safety		Research and Innovation	
Education, Training, Youth		Foreign and Security Policy		Space	
Employment and Social Affairs		Fraud		Taxation	
		Humanitarian aid		Transport	

13c If Yes, at which level?

○ Local / regional levels
X National level
○ European level
X International level

H Use and dissemination

14. How many Articles were published/accepted for publication in peer-reviewed journals?	***0***

To how many of these is open access[7] provided?	
How many of these are published in open access journals?	
How many of these are published in open repositories?	
To how many of these is open access not provided?	
Please check all applicable reasons for not providing open access:	
❑ publisher's licensing agreement would not permit publishing in a repository ❑ no suitable repository available ❑ no suitable open access journal available ❑ no funds available to publish in an open access journal ❑ lack of time and resources ❑ lack of information on open access ❑ other[8]: ……	

15. How many new patent applications ('priority filings') have been made? *("Technologically unique": multiple applications for the same invention in different jurisdictions should be counted as just one application of grant).*		none
16. Indicate how many of the following Intellectual Property Rights were applied for (give number in each box).	Trademark	none
	Registered design	none
	Other	none
17. How many spin-off companies were created / are planned as a direct result of the project?		none
Indicate the approximate number of additional jobs in these companies:		

18. Please indicate whether your project has a potential impact on employment, in comparison with the situation before your project:

❑	Increase in employment, or	❑	In small & medium-sized enterprises
❑	Safeguard employment, or	❑	In large companies
❑	Decrease in employment,	❑	None of the above / not relevant to the project
X	Difficult to estimate / not possible to quantify		

19. For your project partnership please estimate the employment effect resulting directly from your participation in Full Time Equivalent (*FTE = one person working fulltime for a year*) jobs:	*Indicate figure:*
Difficult to estimate / not possible to quantify	X

[7] Open Access is defined as free of charge access for anyone via Internet.
[8] For instance: classification for security project.

I Media and Communication to the general public

20. As part of the project, were any of the beneficiaries professionals in communication or media relations?

○ Yes X No

21. As part of the project, have any beneficiaries received professional media / communication training / advice to improve communication with the general public?

○ Yes X No

22 Which of the following have been used to communicate information about your project to the general public, or have resulted from your project?

X	Press Release	X	Coverage in specialist press
X	Media briefing	❑	Coverage in general (non-specialist) press
X	TV coverage / report	X	Coverage in national press
❑	Radio coverage / report	X	Coverage in international press
X	Brochures /posters / flyers	X	Website for the general public / internet
❑	DVD /Film /Multimedia	X	Event targeting general public (festival, conference, exhibition, science café)

23 In which languages are the information products for the general public produced?

X	Language of the coordinator	X	English
X	Other language(s) Arabic		