Building Annotated Written and Spoken Arabic LR’s in NEMLAR Project

M. Yaseen¹, M. Attia², B. Maegaard⁴, K. Choukri³, N. Paulsson³, S. Haamid², S. Krauwer⁵, C. Bendahan³, H. Fersoe⁴, M. Rashwan², B. Haddad⁶, C. Mukbel⁷, A. Mouradi⁸, A. Al-Kufaishi⁴, M. Shahin², N. Chenfour⁹, A. Ragheb²

¹Amman University; AU, Jordan. mYaseen@ammanu.edu.jo
²The Engineering Company for the Development of Computer Systems; RDI, Egypt {m_Atteya, Salah, Mohsen_Rashwan, Mostafa_Shahin, Ragheb}@RDI-eg.com
³Evaluation and Language resources Distribution Agency; ELDA, France {Choukri, Paulsson, Chomicha}@ELDA.org
⁴Center for Sprogteknologi; CST, University of Copenhagen, Denmark {Bente, Hanne, kuadil}@cst.dk
⁵ELSNET, University of Utricht, The Netherlands Steven.Krauwer@ELSNET.org
⁶University of Petra, Jordan Haddad@uop.edu.jo
⁷University of Balamand; UoB, Lebanon Chafic.Mokbel@balamand.edu.lb
⁸Univ. of Mohammed V Souissi –Ecole Nationale Supérieure d’informatique et d’analyse des Systèmes; ENSIAS, Morocco Mouradi@ensias.ma
⁹Faculté des Sciences Dhar El Mahraz, Fès –Département de Math. et Informatique, Morocco nChenfour@fsdmfes.ac.ma

Abstract

The NEMLAR project: Network for Euro-Mediterranean Language Resource and human language technology development and support; (www.nemlar.org) is a project supported by the EC with partners from Europe and the Middle East; whose objective is to build a network of specialized partners to promote and support the development of Arabic Language Resources in the Mediterranean region. The project focused on identifying the state of the art of LRs in the region, assessing priority requirements through consultations with language industry and communication players, and establishing a protocol for developing and identifying a Basic Language Resource Kit (BLARK) for Arabic, and to assess first priority requirements. The BLARK is defined as the minimal set of language resources that is necessary to do any pre-competitive research and education, in addition to the development of crucial components for any future NLP industry.

Following the identification of high priority resources the NEMLAR partners agreed to focus on, and produce three main resources, which are: 1) Annotated Arabic written corpus of about 500 K words, 2) Arabic speech corpus for TTS applications of 2x5 hours, and 3) Arabic broadcast news speech corpus of 40 hours Modern Standard Arabic.

For each of the resources underlying linguistic models and assumptions of the corpus, technical specifications, methodologies for the collection and building of the resources, validation and verification mechanisms were put and applied for the three LRs.

1. Introduction & Background

Human Language Technologies (HLT) are centrally used in modern information technologies allowing humans to interact with computers in a more natural way and in their own language and facilitating the human-machine-human communication. For business as well as for government administration and for the everyday tasks, it is important to be able to produce text efficiently, to translate, to retrieve information, both in written and spoken form. Therefore, to have a full access to the information technologies in different parts of the world it is crucial to build HLT resources in the used languages.

Moreover, the availability of adequate LRs for as many languages as possible and, in particular, of multilingual LRs, is a pre-requisite for the development of a truly multilingual Information Society. By creating a network of qualified Euro-Mediterranean partners NEMLAR succeeded to specify and support the development of high priority Language Resources (LRs) for Arabic in a systematic, standards-driven, collaborative learning context Maegaard et al (2002). The project focuses on identifying the state of the art of LRs in the region, assessing priority requirements through consultations with language industry and communication players, and establishing a protocol for developing a Basic Language Resource Kit (BLARK), which constitutes a must for each and all languages, to allow for automatic processing of the language for the major forms of the region’s predominant language – the Arabic. ELRA and ELSNET have been promoting this concept for different European languages. This concept proved to be very helpful for the Arabic language with its variety of forms Krauwer et al. (2004).

NEMLAR has identified the BLARK for Arabic and it assessed first priority requirements, a review of existing resources that would match the defined requirements and priorities has taken place. These resources have been identified through a survey; the list of resources was tested for fitness for the specified purposes. The project worked out a clear roadmap to fulfill the requirements identified, and to provide and build many of the missing resources.

Language Resources (LRs) are recognised as a central component of the linguistic infrastructure, necessary for the development of Human Language Technologies (HLT), and therefore for industrial development. Other
purposes may be served by the availability of LR{s} such as content industry, cultural heritage safeguarding, etc.

The issue of HLT based on and/or devoted to the Arabic language is now getting prominent; the lack, on the one hand, of resources, and, on the other hand, of real-world applications, highlights the need for improving R&D in this area and for promoting the use of HLT among the potential partners.

Many applications were identified as potential candidates that need annotated corpora, among those are:

- Morphological Analysis and Composition (inflection, derivation, stemming, ...etc.), PoS disambiguator/ tagger, Diacritizer, Sentence Boundary Detection (punctuation), Statistical approach in Named Entity Recognition, Word Sense Disambiguation, Term Extraction, Statistical Approach in Shallow Parsing Applications, Potentially could be used for Syntactic analysis but will be useful if the Statistical approach is used, Sentence Synthesis and Generation, and Proper Names Recognition, Speech Recognition, Broadcast News Transcription.

Following the work carried out within the NEMLAR project it was agreed to focus on three main resources:

- Written corpus: 500K words of annotated, including vowelized corpus taken from news sites, covering a variety of categories, fully packaged both the annotated and unannotated parts.

- Speech corpus for TTS applications: Recordings/ segmentation/ etc. of 10 hours (2x5 hours) for speech synthesis, 5 hours of a male speaker and 5 hours of a female speaker.

- Broadcast news: The data consist of about 40 hours of Arabic data (mainly Standard Arabic from a number of broadcast companies); Transcriptions followed the Transcribers conventions as used by ELDA and focused on the orthographic, named entities, speaker/turn segmentation levels, with no phonetic transcription/ segmentation.

The choice and parameters of the three corpora were motivated by the urgent needs for such resources for research and development in the Arabic HLT and the available resources within the project to accomplish the building task.

For each of the resources, the underlying linguistic models and assumptions are defined and the structure of the produced LR is dissected. Next, the production process covering all aspects especially the source text or speech acquisition, the deployed annotation tools, the work teams and workload distribution among the concerned partners was set up, coordination procedures were established to harmonize the work under consideration and to guarantee consistency, and revision processes for all details were carried out. The validation criteria, process, and results for all of the three resources are then manifested.

2. Annotated Arabic Written Corpus

The Arabic written corpus (WC) was implemented by a joint team of The Engineering Company for the Development of Computer Systems (RDI) in Egypt, and Amman University (AU) in Jordan. Data validation was conducted by ELDA in France with the support of CST, University of Copenhagen Denmark, Utrecht University (Netherlands) and University of Balamand (UoB) in Lebanon. The raw textual data was obtained from Media International the operator of the famous web portal www.IslamOnLine.net (75%), RDI internal documents (20%), Annahar Lebanese news paper (3%), and other free sources (2%). All sources generously provided IPR clearance letters to the consortium. The Arabic NLP visual annotation tool, along with training and final packaging was provided by RDI.

2.1. Corpus Design

The corpus design of the resource is a function of three criteria; sampling strategy, definition of the annotation types, and size.

Sampling parameters that were taken into considerations are:

- Time span (mostly recent; i.e. late 1990’s till 2005)
- Only Standard Arabic is considered as it is the most commonly accepted variant throughout the native Arabic speakers, and also due to its regularity that can be consistently modelled by the available tools.
- Miscellaneous domains (political, scientific ...) are represented according to their importance weights in potential applications.

The following table represents the size of each of the categories in the selected corpora:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Size</th>
<th>% of total size</th>
</tr>
</thead>
<tbody>
<tr>
<td>General news</td>
<td>100,000</td>
<td>20.0%</td>
</tr>
<tr>
<td>Dictionary entries explanation</td>
<td>52,000</td>
<td>10.4%</td>
</tr>
<tr>
<td>Political news</td>
<td>51,000</td>
<td>10.2%</td>
</tr>
<tr>
<td>Scientific press</td>
<td>50,000</td>
<td>10.0%</td>
</tr>
<tr>
<td>Sports press</td>
<td>50,000</td>
<td>10.0%</td>
</tr>
<tr>
<td>Interviews</td>
<td>49,000</td>
<td>9.8%</td>
</tr>
<tr>
<td>Political debate</td>
<td>35,000</td>
<td>7.0%</td>
</tr>
<tr>
<td>Arabic literature</td>
<td>31,000</td>
<td>6.2%</td>
</tr>
<tr>
<td>Islamic topics</td>
<td>29,000</td>
<td>5.8%</td>
</tr>
<tr>
<td>IT Business &amp; management</td>
<td>20,000</td>
<td>4.0%</td>
</tr>
<tr>
<td>Legal domain text</td>
<td>20,000</td>
<td>4.0%</td>
</tr>
<tr>
<td>Text taken from Broadcast News</td>
<td>8,500</td>
<td>1.7%</td>
</tr>
<tr>
<td>Phrases of common words</td>
<td>5,500</td>
<td>1.1%</td>
</tr>
<tr>
<td>Total:</td>
<td>500K words</td>
<td>100%</td>
</tr>
</tbody>
</table>

In addition to the categorization of un-annotated text; these criteria led to the following types of text annotation:

- Arabic lexical analysis.
- Arabic Part-of-Speech (PoS) tagging.
- Arabic phonetic transcription (i.e. vowelization).

These annotation types applied in this corpus were determined in light of:

- The availability of reliable annotation tools.
- The availability of know-how as well as the manpower for the annotation process.
- The preference of fundamental annotations to higher level ones which fits the BLARK concept.

The third parameter of size was mainly constrained by the budget and time allocated for this task in the project. It corresponded to about 25 man-months for the development of this language resource. With the production process described in the next sub-section, these 25 man-months limited the corpus size to 500,000 words. Given more budget in the future, the corpus can be easily expanded to much more size.
2.2. Corpus Production Process

The production process goes briefly as follows: Arabic linguists run a visual Arabic annotation tool provided by RDI; called Fassieh© where they initiate a book project and feed to it their plain text files where each file is considered a logical page. Then, an NLP engine is run automatically on the text of each page with an accuracy of around 93%.

To raise the accuracy to 99%, trained linguists use the visual revision mode of Fassieh© where all the possible analyses of each word are displayed in ranking order, the linguist then has to either approve the 1st most likely analysis (in most of the cases) or might select another one from the list if it is more suitable and fits better, this happens in the about 7% of the cases.

2.3. Corpus description

After the completion of the initial process described above, Fassieh© tool produces four different output files generated from the initial text files, all are in Unicode format text to guarantee maximum portability. The generated from the initial text files, all are in Unicode format.

2.3.1. Lexical analysis files

Each file of this kind is simply the same as the input file with each Arabic word is replaced by its lexical analysis in the following BNF format:

\[
\{W_v; T \} \quad \text{where;}
\]

\[
W_v \quad \text{is the vowelized full form word}
\]

\[
T \quad \text{is the PoS tag mnemonic of the full form word defined as } T_p := T
\]

\[
T \quad \text{is a PoS tag mnemonic}
\]

For the details of the underlying Arabic PoS tagging model see Attia (2005), and Attia and Rashwan (2004).

The PoS tagging corresponding to the same sample fragment input text presented earlier is shown below:

\[
\{ Tp \} \quad \text{where;}
\]

\[
T_p \quad \text{is the PoS tags vector of the full form word}
\]

\[
T \quad \text{is a PoS tag mnemonic}
\]

2.3.2. PoS tagging files

Each file of this kind is simply the same as the input file with each Arabic word replaced by its PoS tags vector in the following BNF format:

\[
\{ (W_v; T) \}
\]

\[
W_v \quad \text{is the vowelized full form word}
\]

\[
T \quad \text{is the PoS tags vector of the full form word defined as } T_p := T
\]

\[
T \quad \text{is a PoS tag mnemonic}
\]

2.3.3. Vowelization files

Each file of this kind is the same as the input file with each Arabic word is fully vowelized (phonetically rewritten) taking into consideration the mutual phonological effects between adjacent words when concatenation speech is uttered.

There’re 4 more vowelization marks that were added to the standard set of Arabic diacritics in order to get a fully vowelized Arabic text that is ready to be 1-to-1 mapped into any phonetic transcription notation (IPA, SAMPA, ...). These symbols are: @ for long vowel, × for unpronounced characters, ^ for unwritten long vowel ALIF, and ~ for ALIF LAYYINA (long vowel ALIF but written as YAA).

For the details of the underlying methodology of the automatic Arabic phonetic transcription see Attia (2005).

The vowelized text corresponding to the same sample fragment input text presented earlier is the following:

\[
\{ wp \} \quad \text{where;}
\]

\[
w_v \quad \text{is the vowelized full form word}
\]

\[
T \quad \text{is a PoS tag mnemonic}
\]

2.3.4. Map file

This file maps each text file to its source (e.g. URL if the source is the www), as well as its domain category.

The detailed specifications document of this corpus can be found in Attia, Yaseen & Choukri (2005).

2.4. Validation of the written corpus

The validation of the written corpus (WC) followed a methodology inspired by and parallel to the methodologies recommended and promoted by ELRA for validation of spoken resources and lexical resources, respectively, Fersøe (2004), Van den Heuvel et al (2002).

2.4.1. The validation criteria

The validation criteria specific to the NEMLAR WC are described in Fersøe & Paulsson (2006).
For the documentation validation, the criteria focus on minimal requirements to the availability and usefulness of three classes of information. Availability means that the information must be included in the documentation, while usefulness means that it must be presented in such a way that a new user may easily access it. This also entails that the documentation must have a manageable size.

The three classes of information are Administrative Information, for instance contact person; owner; producer; distributor; IPR/copyright statement; etc., Technical Information, for instance read-me file; structure, naming and size of discs, directories and files; format of data and annotation files; associated tools; etc., and Content Information, which is corpus specific.

For the formal validation, the criteria address whether the corpus package as such complies with the technical specifications made in the documentation. Formal validation is made on the entire dataset.

For the content validation, the criteria address both the compliance with the documentation and the linguistic correctness of the annotations. The content checks were made manually on samples of approximately 5,000 words. The criteria were Fragmentation and Integrity of the Material (number of fragmented phrases, number of phrases containing offending material), Lexical Analysis (correctness and consistency of lexical analysis), Vowelization (rate and accuracy of vowelization), and PoS Tagging (correctness and completeness in assignment of PoS tags).

### 2.4.2. The validation results

The validation report recommended to the producer to make some easy improvements in the documentation.

As a result of the formal validation some technical repairs were recommended. Some suggested repairs are mainly related to the formatting of the corpus into four different sets of parallel files with different annotations (raw text, lexically analysed text, vowelized text and PoS tagged text). Other suggested repairs relate to the definition and use of non-Arabic alphabetical strings.

The result of the content validation can be summarized as follows:

- No fragmented or offending phrases.
- 0.55% errors in lexical analysis and very few inconsistencies.
- 53 Arabic words not fully vowelized. Inconsistencies in vowelization between the four parallel datasets.
- Some minor errors in PoS-tagging.

### 3. Arabic TTS Speech Corpus

The recordings and transcriptions of the Arabic TTS Speech Corpus were carried out by RDI of Egypt and ENSIAS of Morocco. Data validation was conducted by ELDA with the support of UoB of Lebanon. The data consists of more than 10 hours of read speech, 5 hours from a female speaker and 5 hours from a male speaker.

#### 3.1. TTS speech corpus specifications

The specification of language resources for speech synthesis has been addressed by various authors. According to these specifications language resources have been built for European languages. The aim of this resource was to come up with specifications on language resources (LR) for speech synthesis based on which LRs in a variety of languages can be produced and is intended to serve as a basis for other projects like ECESS (www.ecess.org) that aims at establishing standards for TTS. As a first attempt for Arabic and in the context of NEMLAR the LR should be suitable to build the most advanced state-of-the-art TTS systems (at least for concatenative speech synthesis).

The creation of voices for TTS systems will be based on read speech. For this issue, text corpora were specified which have to be read by two selected speakers.

The main issue in synthesizing speech is to achieve a good coverage on speech segments used in a given language. In order to achieve more or less ‘perfect’ coverage on a variety of different domains a sub corpus called ‘frequent used phrases’ was specified. Linguistic structures found in text (e.g. as found in a newspaper) differ from those found in speech. For this purpose text derived from ‘written text’ and text derived from ‘transcribed speech’ - i.e. from utterances of speech corpora converted to text – were also included. To increase the prosodic coverage of the segments with respect to their position at the beginning and the end of a sentence, a corpus on written text containing many short sentences was also included in the corpus.

The NEMLAR TTS speech corpus is composed of the following three sub-corpora:

<table>
<thead>
<tr>
<th>Sub-corpus</th>
<th>Tokens/Speaker</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1_T: transcribed speech</td>
<td>6600</td>
<td>1.0</td>
</tr>
<tr>
<td>C2_T: written text</td>
<td>16500</td>
<td>2.5</td>
</tr>
<tr>
<td>C3_T: constructed phrases;</td>
<td>10100</td>
<td>1.5</td>
</tr>
<tr>
<td>consists of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3.1_T: frequent phrases</td>
<td>3500</td>
<td></td>
</tr>
<tr>
<td>C3.2_T: missing &amp; rare diphones</td>
<td>6600</td>
<td></td>
</tr>
</tbody>
</table>

The selection of the speakers was done very carefully. Selection criteria are pleasantness of the voice and the suitability for speech synthesis based on concatenation and pitch synchronous manipulation. One male and one female native professional speaker with an age between 22 and 50 were selected. Each speaker recorded the full corpus of 5 hours.

Ideally the recordings should cover different speaking modes and the speech segments should cover all phonetic variations as well as all prosodic variations and all kinds of speaking modes.

All the recordings were made in a studio and comprise two synchronized channels with speech and laryngograph signals respectively. In addition the speech signals had to fulfill the following criteria:

- 96kHz sampling rate
- 24 bit precision
- SNR > 40 dBA
- RT60 < 0.3 s
- Bandwidth: at least 40 – 20’000 Hz

Furthermore, annotation is based on those rules:

- All speech recordings are transliterated in normalized text form using Arabic vowelized scripts.
- All speech transcriptions are tagged (PoS) and annotated with specific markers (e.g. noise, unintelligible words, etc.).
• All speech recordings have to be marked prosodically.
• For the baseline voices, speech recordings are phonetically transcribed, pitch marked and checked manually.

For the detailed specifications document of this corpus see Haamid, Paulsson, Choukri, and Shahin, (2005).

3.2. Validation of the TTS Speech corpus

The validation of the TTS speech corpus follows the standards and guidelines set out by ELRA for Speech Resources validation, Van den Heuvel et al (2002).

3.2.1. The validation criteria

The validation criteria specific to the NEMLAR TTS speech corpus are described in Paulsson (2006-a).

The validation was carried out in two stages: pre-validation and full validation. The pre-validation was intended to provide the producer with some quick feedback of the most evident deviations.

The full validation was carried out once the final resource had been made available. This stage treats all the checks mentioned in the validation specifications and comprises two main parts: a formal validation of the full corpus to ensure that it complies with the specifications and a content validation to check manually a sample of the recordings and transcriptions.

The formal validation includes checks on documentation: contact person, owner, producer, distributor, IPR/copyright statement, structure, naming and size of discs; as well as checks on the database structure: directories and files, format of data and annotation files, associated tools; etc.

The content validation comprises checks of the acoustic quality of the recordings and the correctness and compliance of the transcriptions. The validation was carried out on a sample of about 3k words and includes checks on the orthographic transcription, segmentation, prosodic transcription and pitch marks.

3.2.2. The validation results

For the validation of both the formal and the content, a number of minor errors were detected and the producer was recommended to update the database.

Some of the deviations detected during the validation are listed below:
• Incomplete documentation.
• Some corrupt files.

The checks on WER, segmentation and pitch marks were all fulfilling the requirements set out in the validation specifications.

4. Arabic Broadcast News Speech Corpus

The audio data of the Arabic Broadcast News Speech Corpus (BNSC) was provided by ELDA of France and the transcriptions and post-processing were carried out by RDI of Egypt. Data validation was conducted by ELDA with the support of UoB of Lebanon and MLTC of Morocco. The data consists of about 40 hours of Arabic data and was provided by ELDA, 209 distinct male speakers and 50 female ones appeared in the raw broadcast news data (mainly Standard Arabic from a number of broadcast companies); Transcriptions follow the Transcriber conventions as used by ELDA and focus on the orthographic, named entities, speaker/tturn and segmentation levels. Also a phonetic lexicon in Arabic SAMPA has been included. The elaboration of the specifications has been inspired by preceding projects like NetDC and Ester. RDI was then in charge of the full production of the whole corpus using the Transcriber 1.4.2 http://www.etca.fr/CTA/gip/Projets/Transcriber/.

The specifications of this corpus were determined by ELDA in collaboration with RDI, see Choukri, Paulsson, Haamid, Shahin (2005).

4.1. Arabic BNSC corpus design

The corpus design of the resource is a function of two criteria:
• a sampling strategy,
• a definition of the size of the resource and of each "session/genre" (if we consider this as important)

Sampling parameters that were used for the design are:
• Any special selection of broadcasting company
• Selection of the channel
• Time span.
• Communication context (news, interviews)
• Language variety (Colloquial vs. Standard, Formal vs. Informal, Country/Region dependent (Egyptian vs. Levantine, Maghreb, etc.))
• Any selection with respect to speaker characteristics, etc.

The news broadcasts were recorded in 16kHz, 16 bit and stored in linear PCM format.

4.2. Validation of the BNSC corpus

The validation of the broadcast news speech corpus (BNSC) follows the standards and guidelines set out by ELRA for Speech Resources Validation, Van den Heuvel et al (2002).

4.2.1. The validation criteria

The validation criteria specific to the NEMLAR BNSC are described in Paulsson (2006-b).

The validation was carried out in two stages: pre-validation and full validation. The pre-validation was intended to provide the producer with some quick feedback of the most evident deviations.

The full validation was carried out once the final resource had been made available. This stage treats all the checks mentioned in the validation specifications and comprises two main parts: a formal validation of the full corpus to ensure that it complies with the specifications and a content validation to check manually a sample of the recordings and transcriptions.

The formal validation includes checks of the acoustic quality of the recordings and the correctness and compliance of the transcriptions. The transcription validation comprises checks of the acoustic quality of the recordings and the correctness and compliance of the transcriptions. The transcription checks were carried out manually on a sample of 6 files with a total of 2 hours of speech. The transcription validation
includes correctness of orthography and also the correctness and completeness of noise markers.

4.2.2. The validation results
For the validation of both the formal and the content, a number of minor errors were detected and the producer was recommended to update the database. Suggestions for improvements included completing the documentation, to add table files for speakers, sessions and topics and to add SAMPA and encoding tables.

Some of the deviations detected during the content validation are listed below:
- Transcription errors including: missing ‘soukoun’ at the end, ‘l’ added to the beginning of some words.
- A few names of speakers have been misspelled.
- Some markers are incorrectly used.

5. Distribution
The Written Corpus WC, consists of more than 500K words in 4 versions: raw text, vowelized text, PoS-tagging and lexical analysis. The TTS Speech Corpus includes more than 5 hours of a female voice and 5 hours of a male voice. The Broadcast News Speech Corpus BNCS, consists of 40 hours of transcribed broadcast news speech from 4 different sources in Modern Standard Arabic. Transcriptions include speaker turns, topics, channel information and a phonetic lexicon in Arabic SAMPA.

The three NEMLAR resources will be packaged and distributed by ELRA through the on-line catalogue: http://www.ELRA.info.

6. Conclusion
The NEMLAR network is established, the BLARK document is defined, and now as a result there is a need for working out a clear roadmap to fulfill such requirements, and to provide and build many of the missing resources. The three different LRs produced at the conclusion of NEMLAR, namely: WC, TTS and BNCS will provide researchers with initial resources that could be expanded and utilized to build upon them some relevant applications.

In order to partly overcome the lack of Arabic Language Resources, the NEMLAR partners would like to ensure that the Arabic language obtains the necessary funds to produce the required resources and tools, and to make them widely available as for many other major languages. A task force was formed and met in Copenhagen; at the conclusion of the meeting, NEMLAR Foundation is launched and came up with a report, NEMLAR (2004), which will be used to raise funds and approach organizations who might be interested in supporting such activities to achieve the desired goals. The NEMLAR Foundation can be contacted through ELDA, CST, AU, RDI, UoB or ELSNET.

7. Acknowledgements
The authors wish to thank the European Commission for the support granted through the INCO-MED programme. The INCO-MED programme has enhanced the development of the cultural dialogue and partnerships across the Mediterranean, as well as the advancement of science to the benefit of all involved parties. It was wise to select language technology as one of the areas to support.

The authors also want to thank all of the project participants, cf. www.NEMLAR.org for their contributions.

8. References
Attia., M., Yaseen., M., Choukri., K. (2005), Specifications of the Arabic Written Corpus produced within the NEMLAR project; www.NEMLAR.org
Maegaard, B. (2004): NEMLAR, Language Technology for Arabic. NEMLAR, Center for Sprögteknologi, University of Copenhagen, Denmark.