Specification of validation criteria

Validation criteria for Nemlar Arabic Written Corpus

Hanne Fersøe (CST)
Niklas Paulsson (ELDA)

2006

The NEMLAR project is supported by the INCO-MED programme

http://www.nemlar.org

Contents

1 Executive summary 4
2 Introduction 4
3 Documentation Validation - Criteria 5
4 Formal Validation - Criteria 6
5 Content Validation - Criteria 6
5.1 Sample sizes 6
5.2 Fragmentation and integrity of the material 7
5.3 Lexical Analysis 7
5.4 Vocalization (Vowelization) 7
5.5 POS Tagging 8
6 References 8

1 Executive summary

This document introduces the Nemlar Written Corpus briefly and specifies the sample sizes, the validation criteria, and the error reporting to be applied in the validation of the corpus.

2 Introduction

The Nemlar Arabic Written Corpus is a corpus of approximately 500,000 words collected from various sources. The texts in the corpus are lexically analyzed, phonetically vowelized, and morphologically POS tagged. The corpus comes from multiple domains with the distribution of domains and words approximately as listed below

Written Corpus domains	Written Corpus, approximate number of words
Political News (63)	48,000
Political Debate (22)	30,000
Islamic text (Preaching and others) (12)	29,000
Phrases of common words (6)	8,500
Broadcast News (4)	5,500
Business (10)	20,000
Arabic literature (24)	30,000
General news (159)	100,000
Interviews (18)	56,000
Scientific press (51)	50,000
Sports press (98)	50,000
Dictionary entries explanation (12)	52,000
Legal domain text	21,000
	Total approximately 500,000 words

The corpus is delivered as four versions of data:

- Raw text. This is the original version of the data as it was extracted from the different sources
- Text with Arabic Lexical Analysis. This contains the lexically analysed data.
- Text with Arabic Diacritization (phonetically vowelized). This is identical to the "Raw text" but with diacritics added.
- Text with Arabic POS tags. This contains the POS tagged data.

The validation will be organized in the following steps:

1. The documentation will be read and validated; see the specification of the criteria in section 3.
2. The corpus package will be received/downloaded and in the process of installing it on the validation center's file system, the automatic checks in the formal validation will be carried out, see the specification of the criteria for formal validation in section 4.
3. The samples will be created according to the specification of sampling in the relevant subsections of section 5.
4. Any other formal validation checks will be carried out.
5. The samples will be organized for validation in a format which allows the validator to make notes of errors and other comments.
6. The content will be validated according to the specification in section 5.

7. The validated samples will be analysed and the result of the validation will be presented in the validation report.

3 Documentation Validation - Criteria

In this part of the validation the documentation is checked and it must fulfill the following minimal set of criteria:

Is there a documentation file? Is it named according to the specifications?

Is the documentation written in English?

Does the documentation contain the following minimal set of administrative information?

- Contact person(s) including
  - name
  - address including e-mail and telephone
  - affiliation, position
- Owner and producer of the data (Nemlar) and distributor of the data (ELDA), including copyright (© Nemlar) and/or IPR statements
- Project name in which data was collected/created
- Distribution media, number of disks

Does the documentation contain the following minimal technical information?

- Contents of each disk – directory and file names
- Layout of the disk file system and directory structure (including nomenclature if relevant)
- Read-me file
- Size of directories and files
- Format of data and annotation files
- Information regarding accompanying tools, if relevant
- Description of mark-up and associated definition files (e.g. DTD)

Does the documentation contain the following content information?

- Database collection strategies employed
- Sources of corpus components
- Number of corpus components
- Size per component
- Total corpus size
- Number of files per domain
- Number of words per file and per domain
- Linguistic information
  - Principles and method used for lexicalization (lexical analysis)
  - Definition of vowelization ‘tags’ used
  - Description of principles for vowelization
  - Definition of POS tag set used
  - Description of principles for the POS tagging, including definitions of tags, combinations of tags, allowed under- or non-specification, multi-word tagging (if relevant)
  - Internal quality assurance procedures used during annotation

4 Formal Validation - Criteria

In this part of the validation the technical deliverable, i.e. the package (the total set of files) and its formal technical qualities are checked. The package must comply with what was stated in the documentation, and it must work, i.e. it must be possible to open the files and gain access to the contents etc.

- Are all the files and directories listed in the documentation included with the correct names?
- Do the layout and the directory structure correspond with the documentation?
- Is there a read-me file with the specified contents?
- Is the size of directories and files in accordance with the documentation?
- Is the number of words per file and domain in accordance with the documentation.
- Do the formal properties of the mark-up comply with the specifications (BNF)?
- Is the format of data and annotation files as specified?, specifically
    - Are all non-Arabic alphabetical strings left unchanged in their original places (as in the raw text files).
    - Is each Arabic alphabetical word replaced by {DiacritizedWord; (TypeMnemonic) TypeCode: (PrefixMnemonic) PrefixCode, (RootMnemonic) RootCode, (PatternMnemonic) PatternCode, (SuffixMnemonic) SuffixCode}
    - Is each Arabic alphabetical word replaced by {(DiacritizedWord); POS_tagMnemonic#}
- Are all the files accessible?

An allowed error rate has not been defined. Identified errors will be counted and reported.

5 Content Validation - Criteria

In this part of the validation a subset of the corpus is validated. In the following sections the sample sizes and the content checks are described in more detail.

5.1 Sample sizes

The content validation will consist of manual checks made on subsets of the corpus. The subsets will be created as balanced samples:

- 1% of the raw text data (approx. 5k words) will form the basis for the validation checks described in section 5.2
- 2% of the “Text with Arabic Lexical Analysis” (approx. 10k words) will form the basis for the validation checks described in section 5.3
- 2% of the “Text with Arabic Diacritization” (approx. 10k words) will form the basis for the validation checks described in section 5.4
- 2% of the “Text with POS tags” (approx. 10k words) will form the basis for the validation checks described in section 5.5

The 2% samples will contain the same words (and only full sentences) selected from the three data versions of lexicalized, vocalized, and POS-tagged data, respectively. The samples will be created by randomly extracting 2.81% of the words from each of the following 6 domains:

Written Corpus domain	Written Corpus size, approximate number of words	Approximate number of words to be validated
Political News	48,000 words	1,350

General news	100,000 words	2,810
Interviews	56,000 words	1,575
Scientific press	50,000 words	1,405
Sports press	50,000 words	1,405
Dictionary entries explanation	52,000 words	1,460
	Total number of words, approx. 356,000	**10,005**

5.2 Fragmentation and integrity of the material

Manual checks will be carried out on the 1% sample to check

- Number of fragmented phrases
- Number of phrases containing offending material (fragments, offensive language, use of other languages like colloquial instead of standard, etc.)

An allowed error rate for acceptance/rejection has not been defined. Identified errors will simply be counted and reported.

5.3 Lexical Analysis

Manual checks will be carried out on the sample to check

- Correctness of lexical analysis
- Consistency of lexical analysis

The goal of the correctness of lexical analysis is 100%. An allowed error rate has not been defined.
The goal of the accuracy is 99%. An allowed error rate has not been defined.

Identified errors will be counted and reported and the error rate will be calculated.

5.4 Vocalization (Vowelization)

Automatic checks will be carried out on the sample to check

- Rate of vocalization meaning to determine the number of consonants/letters that have, incorrectly, not been marked with diacritics. The rate of vocalized consonants/letters can be calculated on the basis of this number.

Manual checks will be carried out on the sample to check

- Accuracy of vocalization meaning correctness according to the principles specified in the documentation.

The goal of the rate of vocalization is 100%. An allowed error rate has not been defined.
The goal of the accuracy is 99%. An allowed error rate has not been defined.

Identified errors will be counted and reported and the actual rate and error rate will be calculated.

5.5 POS Tagging

Manual checks will be carried out on the sample to check:

- Correctness in assignment of POS tags.
- Completeness in assignment of POS tags.

The goal of the correctness of POS-tagging is 100%. An allowed error rate has not been defined.
The goal of the completeness is 99%. An allowed error rate has not been defined.

Identified errors will be counted and reported and the error rate will be calculated.

6 References

[1] Choukri, K., Atiyya, M., Yaseen, M., *Specifications of the LRs to be produced within NEMLAR Arabic Written Corpus*. Nemlar Technical Report. 2005.
[2] van den Heuvel, H.*: Validation criteria*. Orientel. Technical Report D6.2. Version 1.2, 2002.