Improving Access to Text
IMPACT

Technical deliverable documentation

D-EE3.4 Core General Lexicon for Dutch – expanded | EE3

Document history

Revisions

Version	Status	Author	Date	Changes
1.0	First version	INL Impact team	22 February 2011	Created
1.1	Draft	"	1 December 2011	Updated version
2.0	Final version	"	20 January 2012	Final approved version

Approvals

Version	Date of approval	Name	Role in project	Signature
1.0	1 March 2011	Katrien Depuydt	Work package Leader	OK
1.0	4 March 2011	Max Kaiser, Hildelies Balk	SP EE leader, Project Director	OK
1.1	1 December 2011	Katrien Depuydt	Work package Leader	OK
2.0	23 March 2011	Max Kaiser	SP EE leader	OK
2.0	23 March 2011	Hildelies Balk	Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
1.0	28 February 2011	Katrien Depuydt	Workpackage Leader
1.0	1 March 2011	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
1.0	7 March 2011	Liina Munari	EC Project Officer
1.1	1 December 2011	Katrien Depuydt	Workpackage Leader
2.0	23 January 2012	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
2.0	6 April 2012	Liina Munari	EC Project Officer

KB IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

Technical documentation: D-EE3.5 Dutch Core Lexicon

1. Partner

INL

2. Deliverable

D-EE3.5 Dutch Core Lexicon – expanded.

3. Background

The lexicon is delivered as an SQL Database and as an XML document according to the IMPACT Lexicon Format. The core general lexicon for Dutch is intended to improve both OCR and retrieval for historical Dutch Documents. The Core General lexicon lexicon for Dutch relies on the following data:

1. The result of dictionary-quotation-based attestation from the Woordenboek der Nederlandsche Taal (WNT)
2. The result of corpus-based lexicon building from a selection of the Dutch "DBNL[1]" historical corpus material
3. The result of corpus-based lexicon building from selections of KB Material (morphological module)
4. Modern Dutch lexicon data from e-LeX and the hitherto unpublished JVK-lex

The material ranges from 1550 to 1970, thus providing a core around which more specific lexicon data based on selected corpora can be developed.

4. Outline of the Functionality

The core general lexicon for Dutch is intended to improve both OCR and retrieval for historical Dutch Documents. Of the functionalities described in the description of the Impact lexicon structure (D-EE.2.1), it supports token-based attestation, dating of attestations, and lemma and main part of speech information for all word forms described in the lexicon.

The IR lexicon currently contains 475498 distinct word forms, 215180 lemmata, 558438 distinct lemma/wordform combinations. Frequency information can be extracted by counting attestations for wordforms and/or lemmata. Although dictionary frequency (number of quotations) is not a complete substitute for corpus frequency, this information can be used to distinguish highly frequent from infrequent words.

[1] www.dbnl.org

The following simple part of speech set is used for the lemma part of speech information:

ADJ	Adjective
ADP	Adposition
ADV	Adverb
ART	Article
CON	Conjunction
INT	Interjection
NOU	Noun
NUM	Numeral
PRN	Pronoun
VRB	Verb
MWE	Multiword expression
PART_MWE	Part of multiword expression
ADV_MWE	Adverbial multiword expression
INT_MWE	Interjective multiword expression
VRB_MWE	Verbal multiword expression
PART_ADV_MWE	Part of adverbial multiword expression
PART_ADP_MWE	Part of adpositional multiword expression

As yet, there are no grammatical features (like tense, number, case, ...) distinguishing the different word forms belong to the paradigm associated with the lemma. In most cases, the field part_of_speech associated with the word form is identical to the one assigned to the lemma.

When a word (for instance a compound) is split in parts in the attestation (for instance “balsem oly”=balsemolie), the two parts are assigned to the complete lemma (balsemolie), are linked by a common group id in the table wordform_groups, and have the part_of_speech field in the analyzed_wordforms table set to “wordPart”. For separable verbs, we distinguish ‘prefixPart” and “mainPart” for this information.

The corpus-based OCR lexicon is based on diplomatically transcribed text of high quality, but contains occasional errors.

The dictionary-based OCR lexicon includes all word forms in the IR lexicon, and an additional 100.000 manually checked word forms from the historical corpus data.

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

Short description of the files composing this deliverable:

File	Description
./Documentation/EE3.5.doc	This file
./IREvaluationSet/DutchIREvaluationSet.sql	Dutch IR evaluation set
./IRLexicon/EE3.5_Dutch_IR.Lexicon.sql.gz	IR lexicon as compressed (gzip) mysql dump
./OCRLexicon/Binary	
./OCRLexicon/Binary/cleanedDutchCorpusBasedDictionary_1.0.dat	Automatically cleaned version of the corpus based Dutch OCR lexicon, binary format for external dictionary interface
./OCRLexicon/Binary/corpusBasedDutchDictionary_1.0.dat	Corpus based Dutch OCR lexicon, binary format for external dictionary interface
./OCRLexicon/Binary/dictionaryBasedDutchDictionary_1.0.dat	Dictionary based Dutch OCR lexicon, binary format for external dictionary interface
./OCRLexicon/ExternalDictionaryInterface/CExternalDictionary.cpp	Example source code for integration with SDK
./OCRLexicon/ExternalDictionaryInterface/CExternalDictionary.h	Example source code for integration with SDK
./OCRLexicon/ExternalDictionaryInterface/charsets.h	Example source code for integration with SDK
./OCRLexicon/ExternalDictionaryInterface/externalDictionary.doc	Documentation external dictionary interface
./OCRLexicon/ExternalDictionaryInterface/ExternalDictionaryDLL.dll	DLL for windows platform containing core functionality for external dictionary implementation
./OCRLexicon/ExternalDictionaryInterface/LanguageConfiguration.cc	Example source code for integration with SDK
./OCRLexicon/ExternalDictionaryInterface/LanguageConfiguration.h	Example source code for integration with SDK
./OCRLexicon/ExternalDictionaryInterface/RecognizerParamsHolder.cpp	Example source code for integration with SDK
./OCRLexicon/SourceMaterial	
./OCRLexicon/SourceMaterial/cleanedDutchCorpusBasedDictionary_1.0.tc	Automatically cleaned version of the corpus based Dutch OCR lexicon, source text file
./OCRLexicon/SourceMaterial/dictionaryBasedDutchDictionary_1.0.type_confidence.txt	Dictionary based Dutch OCR lexicon, source file
./OCRLexicon/SourceMaterial/dictionaryBasedDutchDictionary_1.0.type_frequency.txt	Dictionary based Dutch OCR lexicon, source file
./OCRLexicon/SourceMaterial/dutchCorpusBasedDictionary_1.0.tf.txt	Corpus based Dutch OCR lexicon, source file
./Spelling/DutchPatterns.txt	Spelling variation patterns for Dutch

5. Evaluation

Both technical and scientific evaluation have been conducted. The results are laid down in the deliverable *D-EE2.8 Use of Computational Lexica for OCR and IR on historical documents – a cross-language perspective.*

6. License and IPR protection

According to agreed terms.