Improving Access to Text
IMPACT

Technical deliverable documentation

D-EE3.4 Core General Lexicon for English | EE3

Document history				
Revisions				
Version	Status	Author	Date	Changes
0.1	Draft	Katrien Depuydt	12 December 2011	Created
1.0	Final	"	20 January 2012	Final version delivered

Approvals

Version	Date of approval	Name	Role in project	Signature
0.1	January 2012	Christoph Ringlstetter, Neil Fitzgerald	Internal reviewers (LMU and BL)	OK
1.0	23 March 2012	Max Kaiser	SP EE leader	OK
1.0	23 March 2012	Hildelies Balk	Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
0.1	12 December 2011	Christoph Ringlstetter, Neil Fitzgerald	Internal reviewers (LMU and BL)
1.0	23 January 2012	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
1.0	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

IMPACT – Technical deliverable documentation

D-EE3.4. Core General Lexicon for English

1. Partner

Instituut voor Nederlandse Lexicologie (INL), Leiden

2. Deliverable

D-EE3.4. Core General Lexicon for English

3. Background

The Impact IR and OCR lexica are based on the dated quotations in the Oxford English Dictionary.
A XML extraction of the quotations per entry has been kindly provided by Oxford University press for use in IMPACT.
In order to build the IR lexicon, the quotations have been processed using INL Attestation Tool in order to extract the word forms corresponding to the dictionary headwords. Quotations from before 1500 or explicitly marked as Old English have been skipped, as have quotations with the attribute “info” set to “intro” or “yes”.
The OCR lexica are based on the extraction of all word forms from quotations dating from three periods: 1580-1720, 1700-1800, and 1750-1920.

4. Outline of functionality

The IR lexicon has 297743 lemmata, 532671 distinct word forms and 874311 wordform-lemma combinations.
The OCR lexica have respectively 406296 (1580-1720), 220044 (1700-1800) and 574444 (1750-1920) entries.

File	*Description*
./IRLexicon/EnglishImpactIRLexicon.v1.sql	Mysql dump of English IR lexicon in IMPACT database format
./IRLexicon/EnglishImpactIRLexicon.v1.xml	TEI P5 export of English IR lexicon database
./IREvaluationSet/EnglishIREvaluationSet.sql	Mysql dump of English IR evaluation set
./IREvaluationSet/EnglishIREvaluationSet.tei.xml	TEI P5 export of English IR evaluation set
./OCRLexicon/1580_1720.tf	Type-frequency list of OED quotations dated from 1580 to 1720
./OCRLexicon/1700_1800.tf	Type-frequency list of OED quotations dated from 1700 to 1800
./OCRLexicon/1750_1920.tf	Type-frequency list of OED quotations dated from 1750 to 1920

5. Evaluation

The technical evaluation of the lexicon has been done by INL. The scientific evaluation is to be found in *Use of computational lexica for OCR and IR on historical documents - a cross-language perspective* (D-EE 2.8).

6. License and IPR protection

The lexicon is owned by Oxford University Press. The OUP signed an agreement with INL for 2011, to use their dictionary data for the IMPACT lexicon. Use of the lexicon after the project will have to be negotiated with OUP.

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