Improving Access to Text
IMPACT

Technical deliverable documentation

D-EE3.9 Core Named Entities Lexicon for English | EE3

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Frank Landsbergen	1 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
0.1	January 2012	Katrien Depuydt	WP leader EE2 and EE3	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	1 December 2011	Christoph Ringlstetter, Neil Fitzgerald	Internal reviewers (LMU and BL)
0.1	1 December 2011	Katrien Depuydt	WP leader EE2 and EE3
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
Core Named Entities Lexicon for English

1. Partner

INL

2. Deliverable

D-EE3.9 Core NE lexicon for English

3. Background

The Core Named Entities Lexicon for English is an elaborate database of enriched historical English locations, person names and organisations. This database can be used as a lexicon for OCR and for query expansion in retrieval.

4. Outline of functionality

The database contains historical English locations, person names and organisations from the period 1742 – 1899 from the following Ground Truth sources:

- A selection of newspaper pages from the JISC-data set
- Book 'Alumni Oxonienses: The Members of the University of Oxford, 1500-1714', by J. Foster
- Book 'Wilson's Mercantile Directory of the World'
- Book 'Cassell's Gazetteer of Great Britain and Ireland
- Book 'Dictionary of National Biography'

An elaborate outline of the database structure can be found in the Lexicon Structure Document (D-EE2.1).

5. Evaluation

In total, the lexicon consists of:

- 255,446 lemmata
- 241,060 wordforms
- 727,386 token_attestations

All locations and organizations are linked to a manually verified modern lemma, and, where applicable, to possible alternative names. For the location lemmata, we have matched the wordform against the location data of the following sources:

- U.S. GeoNames database [http://geonames.nga.mil/ggmagaz/]
- NACO authority files [http://www.loc.gov/catdir/pcc/naco/naco.html]
- Wikipedia [http://en.wikipedia.org]

For locations, there are three possible types of variants. First, synonyms and variants from the source data, e.g. 'Adrian's Wall', 'Hadrian's Wall', have been added as verified variants in the table 'ne_variant_relations'. These locations are also linked through a similar persistent_id in the table 'lemmata'. Second, while lemmatizing the locations, we found that Wikipedia often provided useful variant information, e.g. 'Westcott Barton, also spelt Wescot Barton or Wescote Barton'. These have been manually verified and added to the table as verified variants as well.

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

Person names have been manually annotated with structural information (e.g. 'givenname', 'surname', etc.) through the table 'ne_part_information'. We have used the following elements, which are put in the table 'ne_part_types':

1 given name	(John, William, Archibald)
2 surname	(Johnson, Williams, Jackson)
3 title	(baron, mr., mrs.)
4 particle	(of, de, to)
5 suffix	(jr, D.D., XVII)

Person name variants (e.g. 'Johan' / 'Johann') were matched using the NE-matcher from the NERT-package. In this process, all surname- and given name-lemmata have been matched to one another. Only variants with a matching score of at least 80 (on a scale of 0 to 100) have been added to the database as unverified variants. As an extra feature, we added these scores to the table in an extra column 'ne_variant_relation_score'.The variants are linked through the table 'ne_variant_relations'. In some of the source data, synonyms were given, and these have been added as variants in this table as well.

A large part of the source data contained external_id's from the British Library, which have been stored as external_id's in the table 'lemmata'. All attestations from similar documents are linked by a shared document_id.

Named entities that consist of multiple words (e.g. Napoleon Bonaparte, The Northern Echo) are stored both as a whole and by their parts, on the level of both wordform and lemma. For wordforms, these parts are linked through the table 'attestation_groups', for lemmata, the parts are linked through the table 'multiword_analysis'.
Locations are given a modern spelling variant as lemma, and historical variant wordforms are therefore linked by a shared lemma. Person names have the same lemma as their wordform, and variants are linked by the table 'ne_variant_relations'.

6. License and IPR protection

The licensing follows the consortium agreement.
Product will be integrated as a linguistic resource into the OCR workflow and can be used in demonstrating retrieval. The lexicon is a product by BL and INL. For INL, if agreement is reached for this with BL, it would mean that the lexicon will be made available to the research community according to the regulations of the Dutch HLT agency, which means that it is freely available for non-commercial use. Special licenses can be agreed on for commercial OCR vendors cooperating with an IMPACT partner.

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