Improving Access to Text
IMPACT

Technical deliverable documentation

EE3.10 Core Named Entities Lexicon for Dutch

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	INL IMPACT team	22 February 2011	Created (D-EE3.7)
0.2	Draft	Frank Landsbergen	1 December 2011	• Improvement of PND-matches • Elimination of erroneous entries
1.0	Final	"	20 January 2012	Final version delivered

Approvals

Version	Date of approval	Name	Role in project	Signature
0.1	22 February 2011	Katrien Depuydt	WP leader EE2 and EE3	OK
0.1	1 March 2011	Christoph Ringlstetter	Internal reviewer (LMU)	OK
0.1	4 March 2011	Max Kaiser, Hildelies Balk	SP EE leader, Project Director	OK
0.2	January 2012	Lotte Wilms, Christoph Ringlstetter	Internal reviewers (KB and LMU)	OK
0.2	January 2012	Katrien Depuydt	WP leader EE2 and EE3	OK
1.0	23 March 2012	Max Kaiser, Hildelies Balk	SP EE leader, Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
0.1	22 February 2011	Katrien Depuydt	WP leader EE2 and EE3
0.1	23 February 2011	Christoph Ringlstetter	Internal reviewer (LMU)
0.1	1 March 2011	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
0.1	7 March 2011	Liina Munari	EC Project Officer
0.2	1 December 2011	Lotte Wilms, Christoph Ringlstetter	Internal reviewers (KB and LMU)
0.2	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 Dutch

1. Partner

INL

2. Deliverable

D-EE3.10 Core Named Entities Lexicon for Dutch - expanded

3. Background

The Core Named Entities Lexicon for Dutch is an elaborate database of enriched historical Dutch 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 Dutch locations, person names and organisations from the period 1750 and 1945 from the following Ground Truth sources:

- Staten-Generaal Digitaal
- Newspaper articles
- Book 'Kort begrip der waereld-historie voor de jeugd' – J.F. Martinet
- DBNL (Digitale Bibliotheek der Nederlandse Letteren)
- Keyed NE-lists

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:

- 191.855 lemmata
- 194.210 wordforms
- 509.562 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/]
- Wikipedia [http://en.wikipedia.org/]

For the structure of the database the structure of the general lexicon is used. Named entities that consist of multiple words (e.g. *Napoleon Bonaparte*, *Kaapverdische Eilanden*) 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'.

Data from both keyed NE-lists and ground truth material has been used. The reference between the NE's in the database and those in the used datasets goes via the field 'token_id' in the table 'token_attestations'. All attestations from similar documents are linked by a shared document_id.

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'.

The structure of all person names can be obtained through the table 'ne_part_information'. In this structure, we have used the following elements, which are put in the table 'ne_part_types':

1 given name	(Jan, Piet, Klaas)
2 surname	(Jansen, Pietersen, Klaassen)
3 title	(baron, mr., mevrouw)
4 particle	(van, de, in, tot)
5 suffix	(jr, XVII, Abzn.)

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'.

Person names were matched against the *Personennamendatei* (PND) from the DNB. If a match was found, the PND-id was added as an external-id in the table 'lemmata'.

Variants

The table 'ne_variant_relations' holds three types of variants. First, synonyms and variants from the source data (marked by shared external id's), e.g. person names 'Petrus Blomevenne', 'Petrus van Leiden' and locations 'Mons', 'Bergen' have been added as verified variants.

Second, while manually lemmatizing the locations, we often came across useful synonym information, e.g. the fact that *Ábo* is a synonym of the Finnish city *Turku*. This information has been added as verified variants as well.
Third, person name variants have been 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. The result of the matching for given names was manually verified and is therefore tagged as 'verified', the surname matching results have not been manually verified.

6. License and IPR protection

The licensing follows the consortium agreement.
The 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 KB and INL. For INL, if agreement is reached for this with KB, 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