Improving Access to Text
IMPACT

DELIVERABLE SUBMISSION SHEET

Koninklijke Bibliotheek
IMPACT Project Office

P.O. Box 90407
2509 LK The
. .

Visiting address:
Prins Willem-Alexanderhof 5,
The Hague

T +31 70 314 0958
impact@kb.nl

www.impact-project.eu

		To
(Project Officer)	:	Liina Munari
		Directorate-General Information Society and Media EUFO L-2920 Luxembourg
		From
Project Acronym	:	IMPACT Project number : 215064
Project Manager	:	Hildelies Balk
Project Coordinator	:	Koninklijke Bibliotheek
		The following deliverable
Deliverable title	:	Core NE lexicon for German - expanded
Deliverable number	:	D-EE3.8 (includes D-EE3.11)
Deliverable date	:	11/11/2011
Partners responsible	:	LMU
Status	:	☒ Public ☐ Restricted ☐ Confidential
		Is now complete

☒ It is available for your inspection

☒ Relevant descriptive documents are attached

	The deliverable is	
☒	a document	
☐	a website	[URL:]
☐	software	[ ]
☐	an event	
☒	other	[Lexicon]

Sent to Project Officer	:	Liina-Maria.Munari@ec.europa.eu
Sent to functional mail box	:	INFSO-ICT-215064@ec.europa.eu
On date	:	23 December 2011

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

Improving Access to Text

IMPACT

D-EE3.8 Core NE Lexicon for German - expanded (includes D-EE3.11)

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Christoph Ringlstetter	20 September 2010	Created
0.2	Draft	Christoph Ringlstetter	15 July 2011	Pre-final version
1.0	Final	Christoph Ringlstetter	11 November 2011	Final version

Approvals

Version	Date of approval	Name	Role in project	Signature
1.0	3 December 2011	Frank Landsbergen	Internal reviewer	OK
1.0	3 December 2011	Michael Kranewitter	Tool patron	OK
1.0	3 December 2011	Tomaž Erjavec	Technical reviewer	OK
1.0	16 December 2011	Max Kaiser	Subproject Leader EE	OK
1.0	22 December 2011	Hildelies Balk	Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
1.0	11 November 2011	Frank Landsbergen, Michael Kranewitter, Tomaž Erjavec	Internal review (see above)
1.0	3 December 2011	Max Kaiser	Subproject Leader EE
1.0	16 December 2011	Hildelies Balk	Project Director
1.0	23 December 2011	Liina Munari	EC Project Officer

Additional files for download

From the URL's provided below please download the following additional files:

- The annotated corpora zip-files
    - http://www.cis.uni-muenchen.de/people/kristof/ne/austrianreichsrat.zip
    - http://www.cis.uni-muenchen.de/people/kristof/ne/austriannewspapercorpus.zip
    - http://www.cis.uni-muenchen.de/people/kristof/ne/germanhistoricalcorpus.zip
    - http://www.cis.uni-muenchen.de/people/kristof/ne/germanhistorical16holdoutdouble.zip
- The deliverable as a lightweight xml file as we used it for the NE classifiers;
    - http://www.cis.uni-muenchen.de/people/kristof/ne/ne_core_german.xml
- The deliverable as an sql-dump according to the IMPACT lexicon format
    - http://www.cis.uni-muenchen.de/people/kristof/ne/ne_core_german.sql
- A flat list we will use for OCR experiments
    - http://www.cis.uni-muenchen.de/people/kristof/ne/ne_core_german_ocr.all.txt
- In addition to that we have integrated the hierarchical NE's keyed by ONB.
    - http://www.cis.uni-muenchen.de/people/kristof/ne/ne_hierarchical_onb.zip

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-EE3.8 German Named Entities Lexicon

1.) Partner

LMU

2.) Deliverable

D-EE3.8 German Named Entities Lexicon – expanded (includes D-EE3.11)

3.) Background

The combined deliverable D-EE3.8/D-EE3.11 NE lexicon for German is delivered as a MySQL database according to the IMPACT NE Lexicon format as defined by INL, as an XML document and as a flat OCR lexicon. The NE lexicon is intended to improve both OCR and retrieval for historical German Documents. The user requirements formulated by WP-OC-3 (Evaluation tools and resources) determine the scope of the NE lexicon. The scope has been further specified in product description EE.3.8. In the working plan a first release of the NE lexicon had been scheduled after M24. Because of delays in the essential resources such as keyed materials and groundtruth corpora for NE building, the deliverable of the core NE lexicon had been delayed (see also product description EE.3.8). The deliveries of the core lexicon and the extended lexicon has been decided by the work package leader together with the project manager to be merged into one deliverable. NE lexicon development, NE recognition and experiments on improvements of OCR of NE's will continue until M48 and will be disseminated as a scientific paper.

The MySql database, dumped into 'ne_core_german.sql', contains a first release of the NE-lexicon for German. This release contains locations (NE_LOC) and person names (NE_PER) from the following data sets:

- German Historical Corpus
- LMU NE resources
- Wikipedia
- Keyed NE lists of the Austrian National Library ONB

The release contains 1.516.278 Named Entities. For the structure of the database the augmented structure of the general lexicon as defined by INL is used. A subset of the locations from the keyed data was supplemented with automatically processed variants which are assumed to be the subsequent columns of an entry tab-separated. The link between variants is organized through the ne_variants_relations table.

For the locations of the ONB data a modern spelling variant was assigned as a lemma to a historical variant through the INL Toolbox for Named Entities variant resolution matching and classification -Deliverable D-EE2.3b. If the confidence value of the matching against a collection of modern place name lexica is not high enough, the variant is not applicable. The link between variants is organized through the shared lemma.

Both person and location names are assumed to have the lemma as their word form if no other information is available.

Since the LMU lexica does not contain complete persons respectively no save information on the composition of names is available, the parts tables of the persons are left empty. Variants of person names are considered only if they have been keyed accordingly, this means if an ONB <PERS> file has several columns. For this release, 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 is established via the field 'token_id' in the table 'token_attestations' where a filename is stored.

Ambiguities

The lexicon has been extended with keyed data which partly was ambiguous with respect to the structure and semantics of the columns and with respect to the language tag. Part of the information was tagged unknown. The focus for further experiments will be laid on OCR improvement specifically for NE's with an inclusion of NE recognition into the IMPACT Postcorrection Tool and with a series of OCR experiments to test the effects of external NE resources on FineReader's performance specifically on NE tokens.

Status

The NE lexicon for German relies on the following data:

1. Keyed Named Entities ONB Data available from June 2011
2. The result of corpus-based NE lexicon building from a selection the German historical corpus as used for EE3 with an extension for the 16th century German that has been available for NE annotation from January 2010.
3. Relevant selections from open source datasets, in particular Wikipedia and geographical Gazetteers
4. CIS NE resources.

The German NE lexicon consists of a core set of named entities which are likely to appear in a wide variety of texts, with extensions specific to text types targeted by IMPACT according to scope information provided by OC3, ONB and BSB. For BSB a focus for the 16th century has been defined and an Early High German corpus has been keyed. For D-EE3.11 a complete NE lexicon for this corpus will be included into the lexicon. For ONB the NE included in the keyed materials have been defined as focus area. This data has been available for LMU from June 2011. Additionally data of the CERL thesaurus has been investigated but since CERL is an existing resource in itself it has not been included into the lexicon (see below).

Delivery Status

Delivery D-EE3.11 is composed of three parts. The core NE lexica for German, a set of NE annotated gold standard resources (corpora) and a series of experiments for NE recognition on historical documents exploring two different approaches, a statistical and a rule based.

Achieved:

a) German NE core lexicon has been built from keyed materials, corpus-based sources and additional public resources. Lists of important NE's, NE contexts such as titles extracted from Wikipedia and other sources have been included in the lexica.
b) The ONB lexicon has been partly enriched with modern lemmas through the INL Toolbox for Named Entities variant resolution, matching and classification -Deliverable D-EE2.3b.
c) An OCR lexicon for the experiments with the ABBYY external dictionary interface has been compiled.
d) Procedure for gold standard annotation defined and data with tools for manual annotation (produced by INL). A sample of the German Main corpus 25% (500,000 tokens) from 100 sources have been annotated

according to IMPACT NE format. ONB Reichsratsprotokolle 31,000 tokens from 81 sources, ONB Newspapers 228,000 tokens from 128 sources.

e) Automatic NE detection (geographical entities <LOC>, persons <PERS> and organizations <ORG>) for German has been implemented with two different approaches: a rule based approach uses tuned local grammars in connection with the collected NE lexica; a supervised machine learning approach (Stanford Parser) was evaluated in different settings with and without lexica, with and without specific training.

4.) Delivery Approach

Output	Delivery method
Annotated Dataset	Zipped utf-8 text directories
Lexicon Data	SQL file, XML file for Human Readers, utf-8 txt list for OCR
Hierarchical Lexicon Data	XML Tree File, 20 utf-8 files with annotation
Recognition Experiments	PDF document

3.8.1 German NE Core Lexicon

The core NE lexica with altogether 1.5 million entries had to be developed using the tools developed in D-EE-2.3 and D-EE-2.4. The selection of corpus material to be processed (ONB) has been determined by ONB using coverage tests corresponding to the requirements of OC1. The gold standard research development and test set has been annotated with the INL NE annotation Tool.

3.8.1.1 ONB lexica.

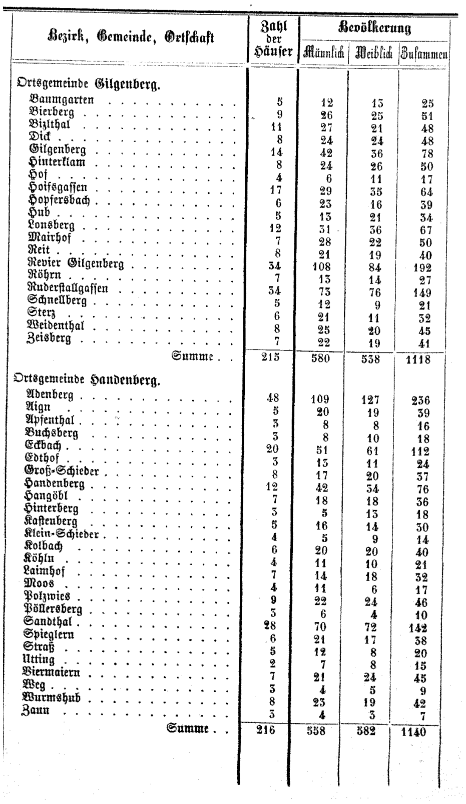
For the ONB lexica 85 works (address registers, place registers…) have been keyed by a service provider. The selection process has been guided by ONB and was monitored by INL with the goal to have the highest possible coverage of Named Entities for 19th century Austrian newspapers under the given budget constraints. For the locations books covering the Habsburg crown lands were chosen. With the budget covered by ONB, and to a smaller part by KB and BL it was decided to key approximately 40,000 pages.

3.8.1.2 CIS lexica.

For the CIS lexica a couple of collected additional lexica with over 10,000 entries from two Master's theses (Ekaterina Chelyapina, Galina Belova) have been used. Additionally Names lists on given and surnames of CIS with over 300,000 entries and a large lexicon of geographical entities with over 300,000 entries (Dr. Sebastian Nagl) have been integrated.

3.8.1.3 ONB hierarchical Lexica.

In addition to the core lexica coded in IMPACT NE format, 20 works with 137,693 hierarchically encoded Named Entities have been rekeyed partially maintaining their hierarchical structure. In the original files a four level hierarchy is coded with numbers at the beginning of each line. The default case is 1 for Bezirkshauptmannschaft/Stadt mit eigenem Statut (City), 2 Gerichtsbezirk (jurisdiction), 3 Gemeinde (town), 4 Ortschaft/Stadtteil (municipal, township).[1] The data is delivered as original files with the hierarchy kept as it is and as an XML Tree. In this XML structure every node has exactly one predecessor. To a specific node successor and predecessor nodes can be retrieved. One obvious application is query expansion. Every NE element has an element "children", including the direct successors.



Bezirk, Gemeinde, Ortſchaft	Zahl der Häuſer	Bevölkerung		
		Männlich	Weiblich	Zuſammen
Ortsgemeinde Gilgenberg.				
Baumgarten	5	12	13	25
Bierberg	9	26	25	51
Bizlthal	11	27	21	48
Dick	8	24	24	48
Gilgenberg	14	42	36	78
Hinterklam	8	24	26	50
Hof	4	6	11	17
Hoifsgaſſen	17	29	35	64
Hopfersbach	6	23	16	39
Hub	5	13	21	34
Lonsberg	12	31	36	67
Mairhof	7	28	22	50
Reit	8	21	19	40
Revier Gilgenberg	34	108	84	192
Röhrn	7	13	14	27
Ruderſtallgaſſen	34	73	76	149
Schnellberg	5	12	9	21
Sterz	6	21	11	32
Weidenthal	8	25	20	45
Zeisberg	7	22	19	41
Summe	215	580	538	1118
Ortsgemeinde Handenberg.				
Adenberg	48	109	127	236
Aign	5	20	19	39
Apfenthal	3	8	8	16
Buchsberg	3	8	10	18
Eckbach	20	51	61	112
Edthof	3	13	11	24
Groß-Schieder	8	17	20	37
Handenberg	12	42	34	76
Hangöbl	7	18	18	36
Hinterberg	3	5	13	18
Kaſtenberg	5	16	14	30
Klein-Schieder	4	5	9	14
Kolbach	6	20	20	40
Köhln	4	11	10	21
Laimhof	7	14	18	32
Moos	4	11	6	17
Polzwies	9	22	24	46
Pöllersberg	3	6	4	10
Sandthal	28	70	72	142
Spieglern	6	21	17	38
Straß	5	12	8	20
Utting	2	7	8	15
Viermaiern	7	21	24	45
Weg	3	4	5	9
Wurmshub	8	23	19	42
Zaun	3	4	3	7
Summe	216	558	582	1140

Example of an image from the original ONB hierarchical files

3.8.2 NE annotated materials – gold standard corpus

Under the lead of the INL Leiden and according to standards outlaid for example in the 1999 Named Entity Task Definition (Chinchor et al.) a set of different NE corpora have been annotated using a NE annotation tool implemented in the EE2 work package by the Leiden team. The annotated corpora are delivered and will be valuable resources for future research on Named Entities in historical documents which is one of the main areas potentially improving user experience in digital library settings presenting such documents to the public.

3.8.2.1 German Main historical corpus NE annotated materials

For the project on the German lexicon in 2008 we started corpus construction searching for appropriate historical

[1] However, these levels are not consistent for all files, e.g. Bosnia: 1 Kreis, 2 Bezirk, 3 Expositur, 4 Gemeinde, 5 Ortschaft. Bukowina: 1 Bezirkshauptmannschaft / Stadt mit eigenem Statut, 2 Gerichtsbezirk, 3 Gemeinde, 5 Gutsgebiet. Croatia and Hungary: 1 Komitat, 2 Stuhlbezirk bzw. Stadtbezirk, 3 unclear, 4 Gemeinde, 5 Ortschaft.

texts in the Internet. We found three interesting collections. The *Bonner Frühneuhochdeutschkorpus* consists of 40 sources from 1350 to 1700 sorted by language regions.[2] Each text contains approximately 30 pages. The *GerManC Corpus* contains 50 newspaper texts from 1650 to 1800 of five regions (northern Germany, western central Germany, eastern central Germany, upper western and upper eastern Germany).[3] Each text represents a sample of 2,000 words. In addition, we manually selected a sample of 53 twice proofread German texts from 1504 to 1904 found on the *Wikisource Project.*[4]

The Institute for German Language (IDS Mannheim) kindly supported our work by providing their *Historisches Korpus IDS* for lexicon building*.* This collection presents the largest part of the first version of our corpus.[5] It contains 408 re-keyed texts of various lengths with a range in time from the year 1700 to 1918, covering different regions and genres such as lexica, newspaper and journal articles, scientific texts, legal texts, literature, and philosophy. The total number of tokens (words in running text) is 3,044,255. Texts whose dates of publication are unknown were excluded from the collection used for lexicon building.

In collaboration with the Bavarian State Library (Bayerische Staatsbibliothek - BSB), we compiled a selection of documents for Early New High German. The *BSB-LMU corpus* consists of 101 works with 1,766 pages, adding up to approximately 858,000 tokens. The materials of the corpus in version 1.0 merged with the new extension establish the current historical corpus for lexicon acquisition in version 2.0. It contains 3,552,690 tokens (words in running text) and 369,730 types (unique words). From these materials a sample of approximately 500,000 tokens has been selected and was manually NE annotated by two linguists.

3.8.2.2 Austrian Reichsrats Protocols NE annotated materials

From the Austrian Reichsrats Protocols collection a sample has been selected to be integrated into the IMPACT groundtruth dataset. From this sample a set of 31,000 tokens from 81 randomly selected pages was prioritized, delivered in 05/2011 and have been annotated for the deliverable.

3.8.2.2 Austrian Newspaper sample annotated materials

From the Austrian Newspaper collection a sample has been selected to be integrated into the IMPACT groundtruth dataset. From this sample a set 228,000 tokens together 128 pages from 10 different newspapers have been NE annotated.

3.8.3 NE Recognition on German documents

For the recognition of NE in German historical documents two different approaches have been tested: a rule based approach relying on local grammars and a statistical approach relying on a classifier implemented by the NER group at the University of Stanford. The statistical classifier has already been used on a demonstrator dataset of ONB in connection with the IR demonstrator engine implemented by INL Leiden and was presented at several demonstration events.

[2] http://www.korpora.org/Fnhd/
[3] http://www.llc.manchester.ac.uk/research/projects/germanc/pilot/,(Scheible et al. 2011)
[4] http://www.wikisource.org
[5] http://www.ids-mannheim.de/ll/HistorischesKorpus

3.8.3.1 Named Entity Recognition on historical texts using local grammars

As a first approach for named entity recognition we were applying local grammars. We created several specialized dictionaries for simple terms as well as for multi-word terms which can deal with a substantial part of the structures marking Named Entities in German historical texts as found in the main German historical corpus used in Impact. The convention of dictionary development according to the DELA format allows for the application of local grammars within the LGPL4 software Unitex5. This platform provides all linguistic tools necessary for the processing of big corpora and enables the efficient handling of electronic lexica. Additionally, the development of local grammars, represented by directed acyclic graph (DAG) structures (cf. Example figures bellow), is supported by a graphical development tool.

The lexica compiled from the Core German NE lexicon have the following structure:

<Matchstring> <classification tag> an example is *Aachner Gymanasium, N+ORG+Cerl.* Unitex works with UTF-16 Little Endian which requires some work on file conversion. The lexicons and the application texts were normalized to lower case since casing is not secure especially for older texts from the 18th 17th and 16th century. Starting points for the local grammars have been two Master's theses at CIS in 2008. The Person lexica have been verified against a large full form lexicon to mark ambiguities that were treated in a separate graph which demanded further evidence to match.

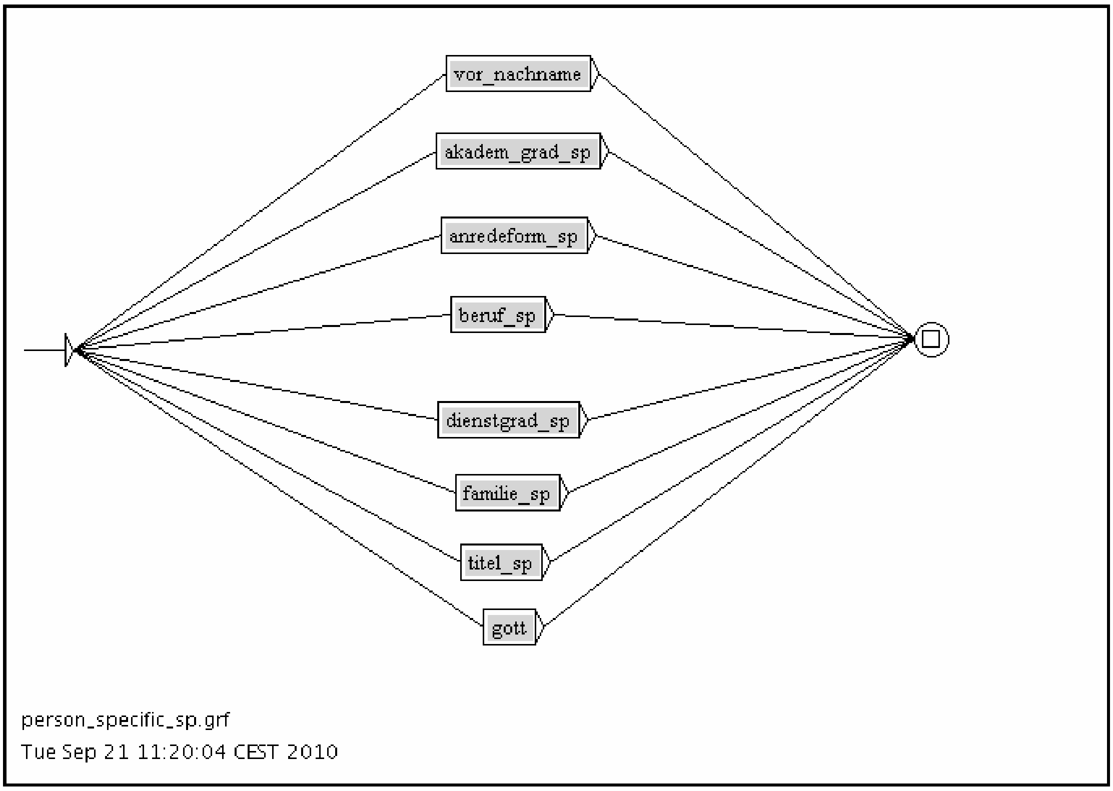
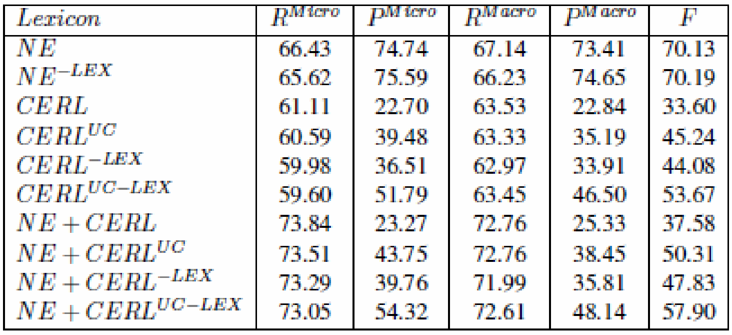


Figure 1. Main graph for person recognition with local grammars pointing to 8 sub graphs.

For the lexica used in the graphs a variety of settings has been conducted leaving for the final experiments an optimal setting of general lexica which then were augmented by the data specific ONB lexica derived from the keyed data. A finding was that the use of big external resources such as the CERL thesaurus led to a drop in performance.



Lexicon	R^{Micro}	P^{Micro}	R^{Macro}	P^{Macro}	F
NE	66.43	74.74	67.14	73.41	70.13
NE^{-LEX}	65.62	75.59	66.23	74.65	70.19
$CERL$	61.11	22.70	63.53	22.84	33.60
$CERL^{UC}$	60.59	39.48	63.33	35.19	45.24
$CERL^{-LEX}$	59.98	36.51	62.97	33.91	44.08
$CERL^{UC-LEX}$	59.60	51.79	63.45	46.50	53.67
$NE+CERL$	73.84	23.27	72.76	25.33	37.58
$NE+CERL^{UC}$	73.51	43.75	72.76	38.45	50.31
$NE+CERL^{-LEX}$	73.29	39.76	71.99	35.81	47.83
$NE+CERL^{UC-LEX}$	73.05	54.32	72.61	48.14	57.90

Table 1: Recall(R), Precision(P), and F measure for Micro and Macro all Named Entities recognised with Local Grammars under different lexicon settings.

3.8.3.2 Named Entity Recognition on historical texts using a statistical classifier

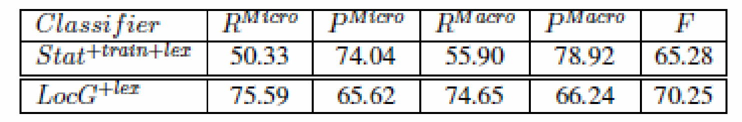
In a pre-test (see the EE2.3 documentation document) an exhaustive test of several statistical NE recognizers has been conducted by the INL group. It showed that the NER Tool provided by the NLP group of the University of Stanford outperformed other implementations significantly. The Stanford Tool is a statistical classifier based on Conditional Random Fields implemented in Java. It labels sequences of words in a text which are the names of things, such as person and location names as well as organization names as found in the IMPACT data. For the model consult, Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370.

In what follows we present results on different corpora and settings of the tool.

As INL we found that the use of the tool does not lead to the expected low quality on the partly very challenging historical texts with many variants and old spelling. The performance is clearly affected by the quality of the training files, which shows e.g. for the classifier trained on the main historical corpus and then used on the Newspaper corpus. So far we did not address variation in historical language with normalization techniques. Experiments of INL on Dutch data showed a gain of 2% in performance for variation reduction.

3.8.3.3 Experiments for Named Entity Recognition on historical texts

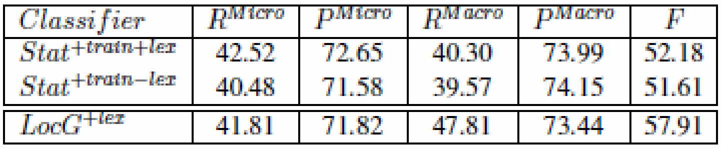
The following experiments try on the one hand to show the overall performance of NE recognition on different corpora of historical documents. On the other hand specifics of the rule based approach (local grammars) and of the statistical approach (CRF, the Stanford tool) shall be compared. The experiments are documented as work in progress and will be available as a research paper and in an updated version for the EE3.11 Extended NE German Lexicon deliverable.



$Classifier$	R^{Micro}	P^{Micro}	R^{Macro}	P^{Macro}	F
$Stat^{+train+lex}$	50.33	74.04	55.90	78.92	65.28
$LocG^{+lex}$	75.59	65.62	74.65	66.24	70.25

Table 2. Experiments on the German General Corpus.

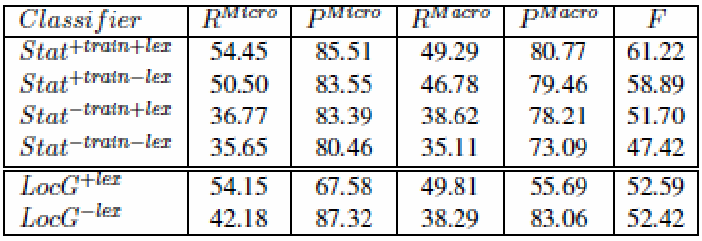
Statistical (Stat) and Local Grammar (LocG) classifiers. Recall (R), Precision (P), (Micro/Macro aggregation) and F measure (Macro) for all Named Entities. Classifiers with general lexica, adapted for the corpus und the Local Grammars developed on the data.



$Classifier$	R^{Micro}	P^{Micro}	R^{Macro}	P^{Macro}	F
$Stat^{+train+lex}$	42.52	72.65	40.30	73.99	52.18
$Stat^{+train-lex}$	40.48	71.58	39.57	74.15	51.61
$LocG^{+lex}$	41.81	71.82	47.81	73.44	57.91

Table 3. Experiments on the German General Corpus – Holdout set for the 16th century.

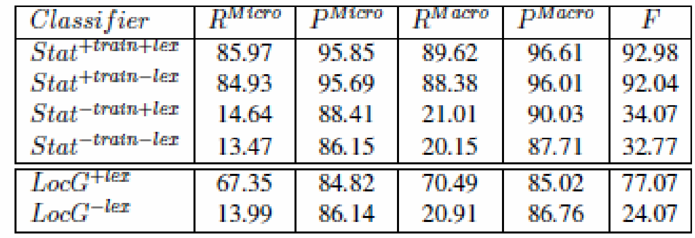
Classifier trained on the German General Corpus, Local Grammars developed on the German General Corpus, holdout set not seen during development. Statistical (Stat) and Local Grammar (LocG) classifiers. Recall (R), Precision (P), (Micro/Macro aggregation) and F measure (Macro) for all Named Entities. Classifiers with general lexica, adapted for the corpus und the Local Grammars developed on the data.



$Classifier$	R^{Micro}	P^{Micro}	R^{Macro}	P^{Macro}	F
$Stat^{+train+lex}$	54.45	85.51	49.29	80.77	61.22
$Stat^{+train-lex}$	50.50	83.55	46.78	79.46	58.89
$Stat^{-train+lex}$	36.77	83.39	38.62	78.21	51.70
$Stat^{-train-lex}$	35.65	80.46	35.11	73.09	47.42
$LocG^{+lex}$	54.15	67.58	49.81	55.69	52.59
$LocG^{-lex}$	42.18	87.32	38.29	83.06	52.42

Table 4. Experiments on the Austrian Newspaper Corpus.

Statistical (Stat) and Local Grammar (LocG) classifiers on the Newspaper Dataset. Recall (R), Precision (P), (Micro/Macro aggregation) and F measure (Macro) for all Named Entities. Classifier with specific lexica (+lex), without specific lexica (-lex) and for the statistical classifier trained on the general corpus (-train) and trained on the Reichsrat Corpus (+train).



$Classifier$	R^{Micro}	P^{Micro}	R^{Macro}	P^{Macro}	F
$Stat^{+train+lex}$	85.97	95.85	89.62	96.61	92.98
$Stat^{+train-lex}$	84.93	95.69	88.38	96.01	92.04
$Stat^{-train+lex}$	14.64	88.41	21.01	90.03	34.07
$Stat^{-train-lex}$	13.47	86.15	20.15	87.71	32.77
$LocG^{+lex}$	67.35	84.82	70.49	85.02	77.07
$LocG^{-lex}$	13.99	86.14	20.91	86.76	24.07

Table 5. Experiments on the Reichsrat Corpus. Statistical (Stat) and Local Grammar (LocG) classifiers on the Reichsrat Dataset. Recall (R), Precision (P), (Micro/Macro aggregation) and F measure (Macro) for all Named Entities. Classifier with specific lexica (+lex), without specific lexica (-lex) and for the statistical classifier trained on the general corpus (-train) and trained on the Reichsrat Corpus (+train).

The results so far show a superior performance of the statistical classifier for specific datasets if it is retrained on an annotated dataset of the same domain. The local grammars seem to be more robust working with the same rules for data of different domains as long as they have the specific lexica available. Without such lexica the performance drops significantly whereas the statistical classifiers show only a very modest improvement for specific lexica on the ONB gold standard sets for Reichsrat and the Newspapers. An interesting observation is that the local grammars suffer sufficiently in Precision if the specific ONB lexica are used which points to a problem with ambiguities. Overall the statistical classifier in the used setting has higher precision combined with lower recall which is inline with the results documented by INL. A final evaluation of the classifiers along with a recommendation for specific situations will be provided with the final deliverable EE3.11.

5.) Form of delivery and license

Form of delivery: cf. above licensed according to consortium agreement.

The product will be integrated as a linguistic resource into the OCR work flow, OCR lexica built from the Core NE lexicon will be integrated on demand into products developed in TR3 and TR5. If budget is available, the recognition pipeline will be integrated into the LMU postcorrection Tool.

6.) Evaluation

The lexicon in SQL and in XML version has been checked for errors. The structure meets requirements formulated in IMPACT deliverable database structure D.EE2.1. Experiments on the delivered gold standard corpora have been conducted.