Technical deliverable documentation

D-EE2.8 Use of Computational Lexica for OCR and IR on historical documents – a cross-language perspective (includes D-EE3.12)

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Initial	Jesse de Does et.al.	21 January 2012	
1.0	Final	Jesse de Does et.al.	1 February 2012	Comments integrated from internal reviewer and language partners

Approvals

Version	Date of approval	Name	Role in project	Signature

Distribution

Version	Date of sending	Name	Role in project
1.0	26-1-2011	Janusz S. Bień (UWAR),	EE3 partner
		Tomaž Erjavec (JSI),	EE3 partner
		Karel Kučera (CUP),	EE3 partner
		Isabel Martinez (UA),	EE3 partner
		Stoyan Mihov (BAS),	EE3 partner
		Gilles Souvay (ATILF)	EE3 partner
	27-1-2011	Rafael Carrasco (UA)	Internal reviewer
2.0	1-2-2011	Janusz S. Bień (UWAR),	EE3 partner
		Tomaž Erjavec (JSI),	EE3 partner
		Karel Kučera (CUP),	EE3 partner

Isabel Martinez (UA),	EE3 partner
Stoyan Mihov (BAS),	EE3 partner
Gilles Souvay (ATILF)	EE3 partner
Rafael Carrasco (UA)	Internal reviewer
Max Kaiser (ONB)	EE subproject leader

Cross-language Perspective on Lexicon Building and Deployment in IMPACT

Jesse de Does and Katrien Depuydt (INL), Klaus Schulz, Annette Gotscharek and Christoph Ringlstetter (LMU), Janusz S. Bień (UWAR), Tomaž Erjavec (JSI), Karel Kučera (CUP), Isabel Martinez (UA), Stoyan Mihov (BAS), Gilles Souvay (ATILF)

Technical deliverable documentation 1
1 Introduction: OCR and IR of historical documents 8
1.1 Accessibility of historical text 8
1.2 Historical language change 9
1.3 Importance of language data and computational techniques 9
2 Introduction: IMPACT approach to lexicon building and deployment and historical language modelling 9
2.1 Modelling historical language variation 10
2.2 Lexicon Building 10
2.2.1 Main road for building lexica for historical language 11
2.2.2 Supervised production of ground truth material for all languages 11
2.3 Lexicon deployment in OCR and post-correction 12
2.3.1 Lexicon deployment in OCR 12
2.3.2 Use of the FineReader external dictionary interface in IMPACT 12
2.3.3 Developing OCR lexica from language resources 13
2.3.4 Post-correction and profiling 14
2.4 Lexicon deployment in IR 15
3 Languages in IMPACT 15
3.1 Bulgarian 15
3.1.1 Period and type of material covered 15
3.1.2 Peculiarities of Bulgarian language and spelling in the period covered 16
3.1.3 Typefaces, scripts, special glyphs 16
3.1.4 Resource situation: dictionaries, corpora, lexica 16
3.2 Czech 16
3.2.1 Period and type of material covered 16
3.2.2 Peculiarities of Czech Language and Spelling in the period covered 17
3.2.3 Typefaces, scripts, special glyphs 17
3.2.4 Resource situation: corpora, lexica 17
3.3 Dutch 18
3.3.1 Period and type of material covered 18
3.3.2 Peculiarities of Dutch Language and Spelling in the period covered 18
3.3.3 Typefaces, scripts, special glyphs 19
3.3.4 Resource situation: corpora, lexica 19
3.4 English 20
3.4.1 Period and type of material covered 20
3.4.2 Peculiarities of English language and spelling in the period covered 20
3.4.3 Typefaces, scripts, special glyphs 20

3.4.4 Resource situation: dictionaries, corpora, lexica 20
3.5 French 20
3.5.1 Period and type of material 21
3.5.2 Peculiarities of language and spelling 21
3.5.3 Typefaces, scripts, special glyphs 22
3.5.4 Resource situation 22
3.6 German 22
3.6.1 Period and type of material 22
3.6.2 Peculiarities of language and spelling 22
3.6.3 Typefaces, scripts, special glyphs 23
3.6.4 Resource situation 23
3.7 Polish 24
3.7.1 Period and type of material 24
3.7.2 Peculiarities of language and spelling 24
3.7.3 Typefaces, scripts, special glyphs 25
3.7.4 Resource situation 28
3.8 Slovene 29
3.8.1 Period and type of material 29
3.8.2 Peculiarities of language and spelling 29
3.8.3 Typefaces, scripts, special glyphs 30
3.8.4 Resource situation 31
3.9 Spanish 32
3.9.1 Period and type of material covered 32
3.9.2 Peculiarities of Spanish language and spelling in the period covered 33
3.9.3 Typefaces, scripts, special glyphs 34
3.9.4 Resource situation: corpora, lexica 34
4 Building the lexica and the spelling variation models 35
4.1 Bulgarian 35
4.1.1 Lexicon building with LeXtractor 35
4.1.2 Lexicon of historical words 36
4.1.3 Spelling variation models 36
4.2 Czech 36
4.3 Dutch 37
4.4 French 37
4.4.1 Before GT 37
4.4.2 GT is available 37
4.4.3 New release 38
4.4.4 GT errors 38

4.5 German ... 38
4.6 Polish ... 39
4.7 Slovene ... 40
4.7.1 Lexicon building with LeXtractor ... 40
4.7.2 Corpus annotation with the CoBaLT lexicon building tool ... 40
4.7.3 Lexicon of historical words ... 41
4.7.4 Spelling variation models ... 42
4.8 Spanish ... 42
4.9 Lexicon building: summary ... 43
5 Deploying OCR lexica ... 44
5.1 Deployment and evaluation of lexica in OCR ... 44
5.1.1 OCR evaluation: use of lexica in OCR ... 44
5.1.2 Conclusions and possibilities for further development ... 61
5.2 Profiler experiments and evaluation for several languages ... 61
6 Deployment and evaluation of IR lexica ... 62
6.1 Deployment of IR lexica ... 62
6.2 Evaluation ... 63
6.2.1 Data used for IR evaluation ... 63
6.2.2 IR evaluation ... 64
6.2.3 Corpus-linguistic evaluation ... 64
6.2.4 IR style evaluation of ranked retrieval by lemma ... 68
6.3 Possibilities for further development ... 69
6.3.1 Making pattern matching more specific ... 69
6.3.2 Using (document or period-specific?) pattern weights ... 70
6.3.3 Corpus-based extension of witnessed historical lexica ... 70
6.3.4 Extending modern lexica ... 70
6.3.5 Taking sentence context into account ... 71
7 Conclusions ... 71
8 References ... 72
9 Appendix: Evaluation data used in this report ... 74
9.1 Bulgarian ... 74
9.2 Czech ... 74
9.3 Dutch ... 74
9.4 French ... 75
9.5 English ... 75
9.6 German ... 77
9.7 Polish ... 77
9.8 Slovene ... 77

9.9 Spanish 78

1 Introduction: OCR and IR of historical documents

1.1 Accessibility of historical text

In the wake of current mass digitization projects in many libraries around the world, huge amounts of historical books and documents are brought to the web. In the scientific community, the difficult metamorphosis from historical paper documents to searchable electronic documents has drawn considerable attention[1]. Though many important projects such as Google Books have already been on their way for some time, from a scientific and technical point of view the path from historical books in paper format to searchable documents in digital libraries is characterized by two core problems that still have not been solved in a fully satisfactory way.

A first difficult step is the conversion from paper to electronic form using optical character recognition (OCR). In spite of recent and ongoing improvements, the quality of OCR'ed historical texts is often still low. This is due to several reasons. Historical fonts often differ per book and are difficult to read. The quality of the paper and the images of historical documents is often suboptimal due to distinct forms of noise and geometric distortion. Furthermore, linguistic components and resources of current OCR systems are often not 'aware' of the kind of language variants found in historical texts. Each input text typically comes with its own specific mixture of features and problems, which explains why the quality of OCR results for historical documents may range from excellent to hardly acceptable.

Even if historical texts are recognized in a perfect way, a second general problem is caused by historical spelling and language variation found in the documents. Most users of digital libraries are not familiar with historical language and want to use modern spelling to search in historical documents. Due to historical language changes and the lack of standardization of orthography in earlier centuries, any modern word may occur in many different spellings in historical documents. When using the current standard techniques in Information Retrieval (IR), these hits are missed, resulting in low recall. To obtain satisfactory answers to search queries, the gap between modern and old spelling needs to be bridged using appropriate methods.

From a cross-language perspective, an additional major challenge in resolving this issue is the fact that for each language the point of departure is different. The level of experience with digitization of historical text material and the amount of reliable text in original historical spelling available is different for different languages (countries). There are also significant differences as to availability of suitable OCR technology (e.g. some Cyrillic characters used in 19th century Bulgarian are not supported by standard OCR engines). For some languages (English, French, German, Spanish) preliminary lexical support for historical language is available in commercial OCR engines, for others this is completely lacking. There are similar differences in the availability of lexical resources for supporting information retrieval on historical documents. Even finer levels of granularity have to be taken into account, because of distinct spellings and alphabet conventions in different periods for different languages.

Language work in IMPACT addresses 9 European languages: Bulgarian, Czech, Dutch, English, French, German, Polish, Slovene, Spanish. We have endeavoured to deal with the aforementioned problems by developing different strategies for lexicon development, taking distinct points of departure into account. These strategies have been implemented in a set of tools to support the construction of historical lexica starting from digital corpora, lexica and electronic historical dictionaries. Using these tools, both OCR and IR lexica have been developed for all 9 languages. In addition to this, a set of tools to exploit the lexical data in both OCR and IR has been developed. To implement lexicon-supported OCR, we have developed a module enabling the use of these lexica with the well-known ABBYY FineReader OCR engine. A retrieval layer on top of the widely-used Lucene search engine has been developed to exploit the historical lexica of all IMPACT languages in IR.

[1] For example: Choudhury et al, 2006, Furrer and Volk 2011.

A problem often underestimated is the question of how to measure progress. In computational linguistics, the importance of good evaluation methods and gold standard data is widely understood, and data sets are available. In digitization, the situation is different. One of the important contributions of IMPACT is the development of an extensive ground truth set in a unified format for 9 languages, both for evaluation of OCR and IR technologies. Apart from the data, a major achievement has been the development of a unified framework for evaluation for 9 languages.

Our final evaluation shows that we have reached significant improvement for all languages. To briefly summarize the outcomes of the evaluation, in OCR we see an improvement of the word recognition rate ranging from 10 to 30%. For IR, we see significant improvement in the recall of historical word forms using modern lemmata as a search key.

The practical progress of the language work in IMPACT is already reflected in take-up of project results in various settings. A commercial provider purchased the Dutch historical OCR lexicon; the Slovene library partner currently deploys the Slovene IR lexicon in its search engine.

Tools and datasets will be made available at www.digitisation.eu, the website of the IMPACT Centre of Competence. A detailed publication of the above mentioned work is envisaged for the first half of 2013.

1.2 Historical language change

Historical language change applies to different aspects of a language. Not only the vocabulary as such changes, with words no longer in existence or not yet existing, but languages also change with regard to grammar, spelling and word segmentation. A typical example of a word not existing in modern language any more is the English pronoun *thou* instead of *you*. Words like *internet* or *computer*, on the other hand, are very recent and will not appear in historical documents. *Ghe-* as in *ghebroken* ('broken') is a typical spelling for historical Dutch, not for modern Dutch, which uses *ge-* (*gebroken)*. In Bulgarian or Dutch, nominal declension still occurs in historical language, but in modern language the case system had almost entirely disappeared. And many historical languages use different word segmentation. A good example of this are the clitics. In historical Slovene, for instance, the negative particle *ne* is often glued to the following word; in Dutch, clitics with pronouns and determiners do occur, but without any form of consistency.

For each language, the degree of language change and the time frame in which it occurs is different. Nineteenth-century English is near to modern, while theDutch spelling has significantly changed since then. Historical lexicon building is only worthwhile if we focus on language periods sufficiently different from the modern language.[2]

1.3 Importance of language data and computational techniques

Lexicon building for OCR and retrieval rests on two pillars: firstly, language data, and secondly, tools for the automatic processing of language material and the manual correction of the processed language data. The language material is needed for lexicon building and for the evaluation of the OCR and needs to be of ground truth quality. In IMPACT, a large ground truth dataset has been created, to be used mainly for OCR evaluation, but in the case of some languages also for lexicon building. The tools in IMPACT were designed for efficient language data processing, and are largely language-independent. An extensive description of the tools and recipes for lexicon building and lexicon application can be found in the *Lexicon Cookbook* (part of D-EE 2.4).

2 Introduction: IMPACT approach to lexicon building and deployment and historical language modelling

[2] See further in chapter 4 and 5.

When designing a lexicon for historical language, certain choices need to be made. As we will briefly explain below, modern and historical language differ in many aspects.

In the context of the IMPACT project, our attention has been focused on the improvement of OCR and Information Retrieval on historical documents. Due to limited resources in terms of time and manpower, and given the practical goals of lexicon work, not all types of information on historical word forms that might appear interesting from the point of view of a historical linguist were taken into account. One good example is part-of-speech information, which in the IMPACT lexica is basically limited to the main part of speech. Lexical descriptions are limited to the intended applications in the field of OCR and IR.

After a brief quick glance at historical language variation, and a description ofthe core structure of our IMPACT lexica, we will summarize how these lexica are built and used.

2.1 Modelling historical language variation

Historical language variation is not restricted to spelling variation. When looking at larger historical periods it becomes obvious that all aspects of a given language are subject to permanent change. Apart from plain spelling, this includes morphology, syntax, semantics, and pragmatics. In practice, these aspects cannot always be fully separated, since even a word that does not change its spelling may obtain a completely distinct meaning. Changes in meaning and syntax as well as pragmatics were not considered in IMPACT, since the focus was on lexicon building applicable to both OCR and IR. Changes in vocabulary, on the other hand,are reflected in the lexicon content.

The standard case of historical spelling variation is the situation where a simple historical word form w_{hist} corresponds to another simple modern word form $\mathrm{w}_{\mathrm{mod}}$.

Often the difference between the modern spelling $\mathrm{w}_{\mathrm{mod}}$ and $\mathrm{w}_{\mathrm{hist}}$ can be described in terms of *patterns* or rewrite rules. For example, the pattern *t*→*th* explains the difference between the modern German word *Turm* (English: tower) and the historical variant *Thurm*. Another historical variant is *Thvrm*, which is derived from *Turm* by applying the two patterns *t*→*th* and *u*→*v*. In the literature, several authors have used pattern-based matching approaches, often adding probabilistic weights to patterns[3]. Pattern-based approaches have also been used in IMPACT as an important ingredient of lexicon building.

When looking at the links between words in old and modern spelling, pattern-based approaches describe the "regular", or "rule-based", part of language variation. In the case of many correctly paired words (w_{hist}, w_{mod}), the nature of this correspondence cannot be captured with a small set of orthographic rules. Although sometimes there is still a clear correspondence on the orthographic level, there are also cases in which the relation with the modern equivalent of a historical word form cannot be captured in simple spelling variation rules. Indeed, some historical words have no modern pendant at all. That is why for IMPACT, we have gone beyond creating spelling variation rules for historical language by also creating historical lexica.

2.2 Lexicon Building

The construction of special lexica for distinct historical European languages has been one of the key activities of IMPACT. In this section we will briefly describe how such lexica were built in IMPACT. For more details on particular approaches followed by IMPACT partners, please refer to chapter 5.

One of the lessons we have learnt is that there are more ways than one; depending on the resources that are initially available for a given language, different strategies are preferable. Still, in the IMPACT project, there does exist a kind of *main road* for lexicon building. It was developed by the two language groups that first entered the project, INL Leiden and CIS (LMU) Munich. Below we shall

[3] Cf. Ernst-Gerlach and Fuhr 2006 and 2007, Kempen et al. 2006, Koolen et al. 2006.

give a detailed picture of this main road, which later was also taken by new language partners in the second phase of IMPACT.

Even before the new language partners had officially entered the project, individual discussions started on the best strategy for building historical lexica for each of the new languages. When looking at the available resources at each institution, the ”main road” turned out to be convenient for most partners. However, some found that specific paths for lexicon building were preferable. Individual strategies for lexicon building used by some of the new language partners will be described further on, in the later parts devoted to the work of single partners. But first we will describe the main road. We will also remark on an important step in lexicon building that affects the work of all partners: the supervised production of ground truth material for each language.

2.2.1 Main road for building lexica for historical language

Possible strategies for lexicon building are always determined by the available *sources* and *methods/tools*. By a source we mean any document that contains historical spellings in their authentic form. We distinguish between (1) sources in paper or image format and (2) sources in electronic symbolic (textual) format. In IMPACT, we have focused on creating lexica from sources of type (2), i.e. electronic textual sources. As one important exception we mention the production of ground truth from images as described below. The reason to give preference to sources in electronic symbolic form is simple. At some point during the lexicon construction, sources of type (1) first have to be converted to a symbolic form. This process is either expensive (manual keying) or error-prone (OCR).

The most important sources of type (2) are *(i)* electronic corpora and *(ii)* digital lexica or dictionaries. It is important to note that many digital lexica are “human-oriented” and address human readers, say, on the internet. This often implies that data are not presented in a way that enables efficient machine access and automated analysis of entries. In contrast, our IMPACT lexica need to be fully formalized, as they must support machine applications such as automated indexing. In IMPACT, both types of sources (i) and (ii) have been heavily used for lexicon building.

Since electronic corpora and digital lexica/dictionaries appeared to be widespread among possible sources for building lexica for historical European languages, work at the beginning of the IMPACT project was focused on the development of tools that directly support the efficient construction of IMPACT lexica from such sources. Tools and recipes are extensively described in the Lexicon Cookbook.

2.2.2 Supervised production of ground truth material for all languages

Though lexicon building work in IMPACT mainly concentrated on existing *electronic textual* sources with authentic language, ground truth texts from images of historical documents were produced for all languages as an additional support. Ground truth material is not only useful for lexicon building, but also for the development and improvement of tools and for the final evaluation steps. As to lexicon building, the main advantage of ground truth material over existing corpus material, it is possible to focus the selection much more on the time period and genres for which the OCR lexicon is needed.

If treated seriously, the production of suitable ground truth texts comes with many difficulties even after the selection of material. In a sense, producing ground truth from historical documents should not be considered as a special kind of manual copying, but rather as a process of translation. The original paper documents often contain special symbols that do not have an obvious representation as a Unicode character. Service providers need advice on how to represent these symbols. For these and similar reasons, resulting from a lack of knowledge of either the font or the language, a permanent communication between service providers and language experts is needed.

2.3 Lexicon deployment in OCR and post-correction

2.3.1 Lexicon deployment in OCR

To evaluate the contribution of special historical lexica, we need to compare the results obtained by using distinct lexica to the results obtained by a state of the art commercial engine. In the IMPACT project, ABBYY FineReader is the key commercial OCR engine to be improved.

Modern OCR systems use a collection of built-in lexica in the background for improving recognition accuracy. This is also true for FineReader. One option for applying the OCR lexica produced in IMPACT could be to extend FineReader (or any other OCR engine) by building in these lexica in the same way lexica for modern languages are implemented by the engine. However, all such dictionaries are built and maintained by ABBYY, and presently there are no externally available tools for building FineReader dictionaries.

Hence, we had to resort to a different approach. The FineReader Engine SDK has an interface for binding so-called “external dictionaries”. This interface has been improved by ABBYY in the course of the project. The INL implemented this interface in order to conduct the evaluation experiments in which OCR lexica with historical vocabulary for all IMPACT languages were used to improve FineReader (section 7).

2.3.2 Use of the FineReader external dictionary interface in IMPACT

Our main aim was to provide an implementation of the FineReader external dictionary interface that would enable IMPACT members to run tests and experiments without being familiar with the exact details of how lexica are used inside FineReader. A usable implementation of this interface requires:

1. Implementation of a C++/COM class interface. Briefly, this consists of implementing two methods:
    a. A method which prunes a “fuzzy set” of word recognition candidates to the subset of linguistically valid ones, providing each valid recognition candidate with a confidence score between 1 and 100.
    b. A method which decides whether a set of recognition candidates contains a prefix which can be extended to a valid word.
2. Development of a simple FineReader SDK-based OCR-executable application which actually uses this implementation during recognition.

The IMPACT implementation consists of:

1. The definition of a “plain C” version of the external dictionary interface, and the development of a Windows DLL implementing this plain C interface, using (a binary compilation of) a static word list with “confidence” information to prune and weigh recognition candidates.
2. An executable which is an adapted version of the CommandLineInterface SDK demonstration program which is part of the FineReader engine distribution. The executable implements the External Dictionary Interface by proxy: the actual work is done in the dynamically loaded DLL module, which is specified on the command line.
3. A small utility program to compile a word list with scores to the required binary format required by 1).

Apart from the evaluation in this paper (section 5), the resulting executable has been deployed within the IMPACT Interoperability Framework[4] as a web service by means of a slight adaptation of the existing command line executable wrapper for the original *CommandLineInterface* program. The DLL has been used by Content Conversion Specialists GmbH to test the effect of the Dutch historical

[4] Cf. (Neudecker et. al. 2011).

IMPACT lexica in an actual OCR workflow by means of integration in the docWorks Large Scale Digitisation Workflow system[5].

2.3.2.1 Long s fix

Our external dictionary implementation contains a workaround for a frequent problem in OCR of historical documents: the recognition of long s vs. f. Even when long s is added to the FineReader character set[6], differentiating the two remains problematic. One of our findings is that it is not always an option to relegate this to the post-correction stage, as the s/f problem may cause the engine to output a completely different recognition candidate, which may be beyond repair by post-correction. For instance Dutch *eerste* (first) is turned into/misrepresented as the dictionary word *eerde (*"honoured"). By having the external dictionary basically "accept" the alternative recognition candidate *eerfte* and correcting it to *eerste* before output, it turns out we can improve recognition.

2.3.3 Developing OCR lexica from language resources

Even if good lexica and ground-truth quality corpora are available, the development of an OCR lexicon is not a completely obvious task. Clearly, a significant degree of coverage on the target material is necessary. On the other hand, including rare and unusual words in the lexicon of accepted words may lead to so-called "dictionary hallucinations", where the OCR engine "corrects" a perfectly valid word to a word form from the lexicon (a "false friend").

It should be clear that the universally optimal OCR lexicon does not exist, except for the ideal "Perfect Lexicon" which contains all words occurring on a certain page and no other words[7]. Of course, we depend to a large degree on the strategy used by the OCR dictionary to weigh dictionary confidence against the results of its internal character classification. Nevertheless, in the concrete situation we are in, working mainly with the FineReader engine, we have found some rules of thumb which can guide OCR lexicon development[8].

2.3.3.1 How many (and which) words should we include?

Again, no universally valid rule can be formulated. However, from our FineReader experience we can say:

- When extracting an OCR lexicon from a large corpus, it does not make sense to include low-frequency words (below frequency 5 or 3)
- Similarly, short words (length 5 and lower) should only be included to the point that a certain degree of corpus coverage, as measured within the set of words of a certain fixed length, has been reached[9]. The reason is that unfrequent short words appear to often contribute more to dictionary hallucinations than to correctly recognized words.
- When a lower frequency word is related to a high frequency word by a frequent OCR confusion, it often improves performance to omit the lower frequency word (this may for instance lead to exclusion of words like *fecond* and *fur* from a French OCR lexicon for documents using long s).

2.3.3.2 Assigning confidence weights to words

In the FineReader engine, a confidence level is assigned both to each dictionary as a whole, and to each individual word. We discussed with ABBYY on the translation of corpus frequencies to confidences ranging from 1 to 100, but no universal rule was given. We ended up choosing an

[5] Presentation at 2011 IMPACT conference, (Gravenhorst, 2011)

[6] It is already included in the default character set for languages with an "Old" dictionary (English, French, German, Spanish). There is an option in the SDK to alter the set of accepted characters.

[7] And of course even this need not be optimal in all situations

[8] Trying to elaborate on these heuristics, and also testing approaches to lexicon building for the open source OCR, tesseract is one of the tasks of the second IMPACT extension.

[9] We used 99% with good results for Dutch and seventeenth-century English and Spanish

approach in which the vocabulary was divided into a few (3 or 4) slices with a different confidence level for each slice S_k, depending on the cumulative frequency achieved by $S_1..S_k$.

2.3.4 Post-correction and profiling

Though modern OCR engines reach impressive results, they are generally based on a limited analysis of the document context due to efficiency constraints.

In this situation, the role of post-correction systems is twofold. Firstly, these tools can try to analyze document and textual context in a more careful and systematic way in order to find better judgments for the correct words at specific places of the document. Secondly, they can be used as an advanced basis for interactive correction of OCR results, pointing to suspicious words, providing alternatives. As a matter of fact, to be of any use, post-correction systems need a strong lexical basis. In IMPACT, a post-correction system was developed at CIS (LMU Munich). It is based on a special technique called *profiling* of OCR-recognized historical documents. In what follows we will first summarize profiling techniques. A more detailed description of profiling can be found in an earlier deliverable (D-TR5.1). Afterwards we will briefly describe the post-correction system.

The profile of an OCR'ed historical text comes with global and local information. The *global profile* provides a list of typical recognition errors in the OCR output with an estimate of the number of occurrences of each type of error, a similar list of typical patterns for historical spelling variation found in the (correct version of) the language of the document with an estimate of the number of occurrences of each type of pattern, and a special kind of language model for the input text. The *local profile* provides a ranked list of individual "interpretations" for each single token of the OCR output. Here, an interpretation of an OCR output token represents a hypothesis on the corresponding correct word in the ground truth and on its modern spelling.

The computation of profiles can be considered as a special form of an expectation maximization procedure in which global and local profiles are improved by using an iterative mutual reinforcement principle. Historical lexica that assign corresponding modern words and pattern derivations to historical words are used to support the initialization of the procedure. Interestingly, at this point IR lexica in the above sense are in fact used for OCR purposes.

The profiler also uses a lexicon for the modern language and a set of patterns as resources.

Obviously, profiles of this form offer many interesting possibilities for adaptation of OCR engines and for post-correction. For example, the list of typical recognition errors with the estimate on the number of occurrences in global profiles, as well as the characterization of particular OCR errors in single tokens derived in local profiles could be used to improve the symbol classifiers of an adaptive OCR engine. In post-correction systems, the same kind of information can be used to point to suspicious recognition results, and to fine-tune the generation and ranking of plausible correction suggestions for ill-formed tokens. The characterization of historical spelling variants found in the document, the text model and other local or global information on the language found in the document could be used to improve the linguistic component of an adaptive OCR. As to adaptive OCR, it is important to note that all information is derived in a fully automated way from the initial OCR output. Hence no manual intervention is needed in the complete workflow from the image to the final recognition result.

The post-correction system designed and implemented by CIS uses the results delivered by the profiler for advanced interactive error detection and error correction. Typical classes of errors can be inspected and corrected in one shot, using lists of concordances with images for visual control.

A first comprehensive description of the profiler was given in IMPACT deliverable D-TR5.1, an evaluation is presented in D-TR5.3. A more detailed portrait of the profiler together with further evaluation results will be presented in a forthcoming research paper. The LMU postcorrection system is described in deliverable D-TR-5.2. After further tests and pilots with the postcorrection tool at distinct IMPACT libraries in 2012 we shall also prepare a research paper on the postcorrection system.

2.4 Lexicon deployment in IR

As we mentioned above, Information Retrieval on OCR'ed historical documents is confronted with two interwoven problems, historical spelling variation and errors introduced by the OCR engine. In this context, the above-mentioned profiling technique is interesting because local profiles for OCR output tokens offer hypotheses on both the correct historical word in the text and its modern equivalent. Once the modern equivalent is known, a final step is lemmatization. We can either use IR lexica of the form built in IMPACT or modern lemmatization procedures to completely close the gap between tokens in the OCR output - i.e. the input tokens for indexing - and modern lemmata used in queries.

To simplify the discussion, let us ignore the problem of OCR errors at this point. We may then directly use IR lexica to link tokens in the text to modern lemmata. As an alternative, a combination of matching and lemmatization techniques can be used, assuming that we have a modern lexicon and an appropriate set of patterns. In earlier experiments for German it was found that matching results are acceptable in the case of documents from the eighteenth and nineteenth century. For earlier centuries, spelling variation comes with many irregular cases, and special lexica are required to achieve correct matching. .

The fact is that none of these techniques works in a perfect way. Correct links may be missed and wrong links may be suggested. Still, experiments in IMPACT show that the above-mentioned techniques substantially improve Information Retrieval on historical document collections[10].

A detailed evaluation of the use of IR lexica developed in IMPACT in an IR scenario can be found in section 8. For demonstration purposes, a specialized Information Retrieval engine based on Lucene has been developed in IMPACT.

3 Languages in IMPACT

In IMPACT, lexicon building has been done for nine different languages. For each language, the point of departure was different. This has primarily to do with the characteristics of each language as such. Historical language change applies to different aspects of a language and differs from language to language. For each language, the degree of language change and the time frame in which it occurs is different. So for IMPACT it was important to analyze the situation per language before deciding which language period and type of material would be covered in the project.

Apart from that, the biggest difference between the languages was the availability of language data for lexicon building. While it was not difficult to obtain ample material for lexicon building, dictionaries and corpus material for Dutch, it was completely the opposite for languages such as Bulgarian and Polish. Their ground truth material had to be produced to obtain material for lexicon building.

This chapter will provide general observations per language about the language, its writing system and the resources used for lexicon building, and describe very briefly the period and type of material covered in the project. The actual lexicon building will be discussed in chapter 5.

3.1 Bulgarian

3.1.1 Period and type of material covered

Two periods have been tackled in IMPACT in distinct ways. For early nineteenth-century material, ABBYY has trained fonts to enable recognition of material in Church Slavonic fonts without diacritics. No lexica have been built for this period.

For the late nineteenth century (1882-1903), books and newspapers have been ground-truthed, and OCR and IR lexica have been built.

[10] Gotscharek et. al. 2009 and 2010.

3.1.2 Peculiarities of Bulgarian language and spelling in the period covered

3.1.2.1 Language

Except for changes in vocabulary, the main change is the loss of the case system.

3.1.2.2 Spelling

The main differences between nineteenth-century and modern spelling are described by the historical spelling variation pattern set delivered. Late nineteenth-century Bulgarian used a character set which is slightly larger than the one currently in use.

The following characters (only lower-case given), which occur regularly in the IMPACT ground truth material, disappeared from modern use:

Unicode	Character	Frequency in lexicon	Modern replacement
0x44b	ы (small letter yeru)	3603	и
0x456	i (ukrainian i)	3486	и
0x463	ѣ (yat)	227524	e or я
0x46b	ѫ (big yus)	47609	ъ or a
0x46d	ѭ (iotified big yus)	1829	я

Other important patterns of change include for instance: disappearance of final ъ and ь; modern Bulgarian spelling does not reflect the assimilation of the last consonant in the prefixes раз- and из-; metathesis in ър/ръ and ъл/лъ groups.

There were various attempts to standardize the spelling: Drinov 1862, Gerov 1895, Ivanchov 1899. Ivanchov's spelling[11] was the first officially adopted spelling. With a brief interlude from 1921 to 1923 (Omarchevski's spelling), it was used until modern spelling was introduced in 1945.

3.1.3 Typefaces, scripts, special glyphs

Early nineteenth-century Bulgarian is typeset in Church Slavonic fonts.

The character set for late nineteenth-century Bulgarian is the modern Cyrillic with the addition of big yus, yat, iotified big yus, Ukrainian ie, yera. FineReader had to be trained additionally to recognize big yus and iotified big yus.

3.1.4 Resource situation: dictionaries, corpora, lexica

Corpora: OCR corpus (Hi-res scanned journals and newspapers (1882-1900) used

Tagged Corpora: none

Historical Lexica: none used

Modern language lexica: yes

Dictionaries: no

3.2 Czech

3.2.1 Period and type of material covered

1800-1900; books, newspapers.

[11] http://ivanchevski.grazhdani.eu/ has a firefox plugin for spellchecking with the Ivanchov spelling.

3.2.2 Peculiarities of Czech Language and Spelling in the period covered

3.2.2.1 Language

Language change:

A limited number of endings was dropped or changed in the nineteenth century, and there were minor changes in several paradigms. On the other hand, enormous changes in technical and literary vocabulary took place, marking the nineteenth century transition of Czech from a language surviving, virtually uncultivated, for two centuries in non-prestigious use to a full-fledged language of culture, science and education of its time.

3.2.2.2 Spelling

The Czech spelling based on the principles established at the turn of the sixteenth century was changed by three major reforms during the first half of the nineteenth century (in 1809, 1843 and 1850). While in handwritten texts a lot of instances of older spelling were found for several decades, in printed texts the reforms came to be implemented with a remarkable consistency within years. Although for some time after 1850 they still showed some ambiguity of spelling,, generally speaking orthography was highly unambiguous in the second half of the nineteenth century and showed no systematic changes for over a century. The following example illustrates some of the more important developments:

before 1810:	*(Stalo se)* ***gj to cyzý winau***	(‘It happened to her by someone else’s fault’)
before 1843:	*(Stalo se)* ***gj to cizj winau***	⇦ Jungmann
before 1849:	*(Stalo se)* ***jí to cizí winau***	
after 1849:	*(Stalo se)* ***jí to cizí vinou***	⇦ Kott

3.2.3 Typefaces, scripts, special glyphs

3.2.3.1 Character sets

- Czech fracture used almost exclusively before 1810
- in isolated early experiments with Latin script prior to 1810, the “long s” (Unicode 017F) is used to mark the letter *s* in other than word-final positions, and a ligature of two long s’s, later on replaced with *š* (Unicode 0161), is also found
- modern Czech alphabet (Latin alphabet with diacritical marks) used in prints from about 1810 on

3.2.3.2 Typefaces

Czech fracture, still a problem for OCR, used almost exclusively before 1809, when it started to be rapidly replaced with Latin script.

3.2.4 Resource situation: corpora, lexica

Corpora: yes (Czech National Corpus; transcribed nineteenth century texts included in its diachronic part)

Tagged Corpora (only the synchronic part of CNC, texts from 1945 to date)

Historical Lexica: none

Modern language lexica: yes

Dictionaries for the nineteenth-century Czech: Josef Dobrovský (1802-1822), Josef Jungmann (1835-1839), F. Š. Kott’s (1878–1893, supplements 1896–1897)

3.3 Dutch

Dutch is a West Germanic language. Most speakers live in the European Union, where it is a first language for about 23 million people and a second language for another 5 million people. In the Middle Ages, regional variation was still very significant. An important step in the process of standardization was the first major Dutch Bible translation of 1618, the language of which was mostly based on the urban dialects from the province of Holland.

3.3.1 Period and type of material covered

Period: 1600-1940; books, newspapers, parliamentary papers.

3.3.2 Peculiarities of Dutch Language and Spelling in the period covered

3.3.2.1 Language

Structurally, Dutch has evolved little since the early seventeenth century. The simplification in the vowel system from Middle Dutch to Early Modern Dutch largely took place before the IMPACT period[12].

The grammar of the written language has been simplified: nominal declension disappeared completely except for the genitive form for proper names and use in collocations.

3.3.2.2 Spelling

There is a significant amount of change in the period covered[13]. The initial system in the period of interest for IMPACT is a "many-to-many" mainly phonetic spelling inherited from the Middle Dutch period. A single phoneme can be spelled in many ways and spelling does not unambiguously reflect pronunciation (esp. for long vowels, diphtongs). Clitic combinations were often written together.

There have been various proposals for the standardization of Dutch spelling. Lambrecht[14] advocated a mainly phonetic spelling (in fact proposing that all dialects should spell according to their pronunciation[15]). Another important work is the *Nederduitse orthographie* by Pontus de Heuiter[16]. Spiegel[17] proposes a unified spelling for the vowel system and advocates the analogical principle[18] which became very important in Dutch spelling. These proposals were never widely adopted in real life; eighteenth-century spelling is still unstandardized and very different from modern spelling.

Standardization occurs in the nineteenth century. The first spelling which was, although not mandatory, officially adopted by schools and government institutions was the one proposed by Matthijs Siegenbeek (1774-1854) in his *Verhandeling over de Nederduitsche spelling ter bevordering van eenparigheid in dezelve (1804)*. The first standard spelling to be universally applied was de Vries and Te Winkel's spelling developed for the WNT (Dictionary of Dutch). It was officially adopted in

[12] http://www.dbnl.org/tekst/bree001hist02_01/bree001hist02_01_0037.php#37

[13] Cf for instance Cf: http://neon.niederlandistik.fu-berlin.de/nl/nedling/taalgeschiedenis/spelling/

[14] Joos Lambrecht, Nederlandsche spellijnghe, uutghesteld by vraghe ende antwoorde (1550), http://www.dbnl.org/tekst/lamb011jfjh01_01/

[15] "Zo waar de Zealāder pronūciëerd Jae/daar en behoard hy in tspellen vā den zelụen woorde den Vlámijngh noch den Brábāter niet te volghen/aldus Ja: of hy moeste oac zoa sprékē". As the Zeelandian pronounces *Jae* / so he should not follow in the spelling of this word the Fleming or the Brabantian / in this way: Ja: or he should speak like that.

[16] http://www.dbnl.org/tekst/heui001nede01_01

[17] Twe-spraack vande Nederduitsche letterkunst , http://www.dbnl.org/titels/titel.php?id=spie001twes01

[18] "Nóchtans wilt niet schicken datmen krapt en klabt zou scryven om dat vant een krabben vant ander klappen komt/ óf ghót ende pód om datmen ghoden ende pótten zeit", it does not befit that one should write *krapt* and *klabt*, as from the first, one has[from the first one has?] *krabben,* and from the second, one has *klappen*, or *ghot* and *pod*, because one says *ghoden* and *potten*. http://www.dbnl.org/tekst/spie001wjhc01_01/spie001wjhc01_01_0008.php

1864 in Flanders and in 1883 in the Netherlands. Its successor was Marchant's spelling (adopted 1946-1947), which is the basis for modern Dutch spelling.

3.3.3 Typefaces, scripts, special glyphs

3.3.3.1 Character sets

The following ligatures are found in the Dutch Ground Truth and listed below with their replacement in OCR evaluation and lexicon lookup. Additionally, long s is replaced with normal s.

Unicode	Character	Frequency	Replacement
e6	Æ	190	ae
133	Ĳ	16378	ij
eada	ſt	33773	st
eba2	ſi	4741	si
eba3	ſl	2644	sl
eba6	ſſ	7227	ss
eba7	ſſi	427	ssi
eec4	ck	1	ck
eec5	ct	1791	ct
eedc	tʒ	1	tz
f50a	ď	42	d'
f51e	ſł	2	s
fb00	ﬀ	2187	ff
fb01	ﬁ	1138	fi
fb02	ﬂ	297	fl
fb03	ﬃ	475	ffi
fb04	ﬄ	1	ffl
fb06	ﬆ	23	st

3.3.3.2 Typefaces

In the seventeenth century, both Gothic and Roman were in wide use, and were often combined. The Gothic script used in the older documents is a problem for most OCR software, including FineReader[19].

From the eighteenth century onwards, mostly roman fonts are used. Their shape does not present any particular problems, apart from the often difficult distinction between long s and f. More problems are caused by the often bad condition of especially the newspaper material.

3.3.4 Resource situation: corpora, lexica

Historical Corpora for the IMPACT period: DBNL[20]

Tagged Corpora: none for the IMPACT period

Historical Lexica: none for the IMPACT period

Modern language lexica: yes. Both e-Lex and the smaller JVKLeX have been used

Historical Dictionaries: WNT (Dictionary of the Dutch Language)[21]. Because modern lemmata have been assigned to the historical forms, this is a good starting point for historical lexicon building.

[19] Recently, improvements for older gothic fonts have been implemented in FineReader 10. They will be tested with Dutch material in the second extension to Impact.

[20] Digitale Bibliotheek voor de Nederlandse Letteren, Digital Library for Dutch Literature, http://www.dbnl.org

3.4 English

3.4.1 Period and type of material covered

1497-1900; books, newspapers, papers.

3.4.2 Peculiarities of English language and spelling in the period covered

"The first English writing system using the Roman alphabet was developed in the 7th century, after St. Augustine brought church Latin to the Saxons in Kent in 597. The language and spelling have both changed a great deal since then. They did not start to resemble current usage until 1348, when a series of plagues helped to end French domination over England and the English language. The system from which current English spelling conventions have developed was the one used by the poet Geoffrey Chaucer, who died in 1400."[22]

English became the official language after the Hundred Years' War with France, around 1430. Unfortunately, most of the writing in English was done by foreign scribes and scholars. As a result, the spelling was full of inconsistencies, and later developments did not bring about any great improvement (see the above website for an elaborate description).

Scholars like D.G. Scragg[23] claim that by the early eighteenth century, the English spelling system had become established. However, this holds only true for printed books. In eighteenth-century letters , for instance, a different spelling system is seen throughout that whole period (Svenja Kranich, *An Introduction to Late Modern English*, p 39). By the end of the eighteenth century, partly under the influence of Samuel Johnson's dictionary (which he started in 1755), correct spelling was being propagated, also with regard to letters. However, a certain degree of inconsistency remained at least until the end of the eighteenth century.

What were the characteristics of English spelling in the eighteenth century? The use of extra initial capitals increased until the middle of the century, but then was abandoned completely by the turn of the century . In the eighteenth century, the long s is still in use, as well as ligatures such as <oe> and <ae>. Other spelling variants are, for example, *oul > ol*; *controul > control*; e$ > 0; *confesse > confess*, *eing$ > ing$*; *takeing > taking*, *'d$ > ed$*; *work'd > worked; ick$ > ic$*; and *musick > music*. Quite a bit of use is made of clipped forms (e.g. thro', ém, or can't, …).(Kranich, p. 39 44).

3.4.3 Typefaces, scripts, special glyphs

Except for one title[24], fonts are roman and close to modern. More problems for OCR are caused by the fact that most of the material consists of poor-quality bitonal scans.

3.4.4 Resource situation: dictionaries, corpora, lexica

Corpora: TCP (closed), ECCO (open)
Tagged Corpora (not so much)
Historical Lexica: none used
Modern language lexica: yes (not used)
Dictionaries: Oxford English Dictionary

3.5 French

French is a Romance language spoken as a first language in Europe (France and parts of Belgium and Switzerland), in northern America (Québec), and in the former French colonies in Africa. French is one of the most studied foreign languages in the world. It is a descendant of the spoken Latin

[21] Cf. http://gtb.inl.nl and http://www.inl.nl/onderzoek-a-onderwijs/lexicologie-a-lexicografie/wnt

[22] http://www.spellingsociety.org/spelling/history

[23] A history of English spelling, 1974, 80

[24] "Prologus Here begynneth the prologue of the storye of Thebes", 1497

language but was influenced by the native Celtic languages in Gaul, and by Germanic languages after the Frankish invasion.

French language can be divided roughly into 5 main periods:

- Old French from 842 (first text in French *Les serments de Strasbourg*) to XIIIth century
- Middle French from XIVth to XVth century
- Renaissance French XVIth and XVIIth century
- Classical French XVIIIth century
- Modern French after XVIIIth century

It should be noted that French spelling was largely established at the end of the classical period. In the nineteenth century the main changes were the substitution of all *oi* digraphs that represented /ε / by *ai*, and analogical spelling of plurals, e.g. *parens* → *parents*.

3.5.1 Period and type of material

French is a language studied intensively in France and other European Countries (not only in French-speaking countries, but also in England or Germany, for example), so the computational resources for the French Language are quite numerous. Among those resources we can quote:

- TLF: *Trésor de la Langue Française*, electronic dictionary for XIXth et XXth French (100,000 entries),
- *Morphalou* a large coverage morphological XML lexicon for French (~540.000 inflected forms) based on TLF entries
- *Frantext* textual database, 5000 texts for all periods of French but mainly for Modern French (larger corpora for French)
- DMF: *Dictionnaire du Moyen Français* (1330-1500), electronic dictionary for Middle French (65,000 entries)
- LGeRM, lemmatiser for Middle French, based on DMF entries with 785,000 inflected forms, and 5,000 variation rules.

These are all reference resources for French Language studies and were designed in ATILF from 1960 to now. Analysing these resources, we noticed the XVIth, XVIIth and XVIIIth century were underrepresented. We decided to focus on the XVIIth century, late Renaissance French. It is an intermediate period between Middle French (covered by LGeRM data) and Modern French (covered by Morphalou data).

3.5.2 Peculiarities of language and spelling

As an intermediate language, late Renaissance French is close to Modern French but preserved Middle French spelling archaisms. Although quite easy to read and understand for a French Native, its spelling of words is often suprising. The sense of the words is usually close to their modern sense. So we decided to use the set of lemmata provided by Morphalou (based on the entries of TLF).

The use of diacritics is quite different. Umlauts are used very frequently; acute, grave and circumflex appear to be quite interchangeable. The etymological origin of ê or é (coming from 'es') is still manifest, both spellings being used, for example in *école* (modern) and *escole*; *êcole* is also a valid possibility. A tilde on a vowel may be used to refer to the combination of a vowel and a nasal consonant. For example *ã* is *an* or *am* (before b, m or p). This is a practice from medieval manuscript writing, preserved in printed texts.

'U' can be used instead of 'V' and vice versa. 'I' can be used instead of 'J'. 'Y' can appear instead of 'I', but also the other way round (for example *hipocrisie*).

Word segmentation is quite modern. The use of "–" to connect two or more words is different from modern French. It doesn't exist in Middle French. *Aussi-tost* is used for *aussitôt*, reminding us that this word comes from the agglutination of two words. For example, the adverb *très* (without grave) may be attached to the following adjectives or adverbs *tres-commun*, *tres-avantageusement*.

Many words preserve the Latin consonant groups: *sainct* for *saint*, *subject* for *sujet*.

3.5.3 Typefaces, scripts, special glyphs

The alphabet is identical to the modern one, except for the use of tilde on vowels, and long s. The diacritics are grave, acute, umlaut, circumflex, cedillas.

There is also one ligature that disappeared in modern French: ct.

3.5.4 Resource situation

All resources used were produced by ATILF, so we could freely use them or adapt them.

3.5.4.1 Historical texts

Besides/Apart from the ground truth we took into account the existing textual database Frantext. We selected all text written between 1600 and 1740 and printed during the same period. The IMPACT development GT contains 670,944 tokens (31,379 types).The Frantext subset has 176 texts from 1600 to 1740. They were partially modernized when keyed, 10-20 years ago, at ATILF. The whole corpus contains 10,684,781 tokens (including punctuation), 132,005 types (including words split by line breaks, words, ground truth errors).

3.5.4.2 Morphalou

To our surprise, we found that not all entries of TLF were included in this resource, which means the IMPACT Lexicon will enrich/correct Morphalou.

3.5.4.3 DMF

DMF was used when there was no modern entry in TLF. As the entries of DMF have been modernized it was quite appropriate to use DMF entries.

DMF is not only an electronic dictionary, it provides tools to help build the glossary of a text. Someone transcribing a medieval manuscript and using TEI format can send its transcription to DMF. Each word is lemmatized and finally all words are sorted by lemma. The result is a draft version of a glossary, which needs to be corrected, disambiguated, etc. In fact it is a tool very close to IMPACT tools in functionalities.

3.5.4.4 LGeRM

LGeRM was designed to lemmatize spelling variation in Middle French. So it was necessary to adapt it for the purpose of IMPACT. To be able to lemmatize a text, LGeRM uses a knowledge basis consisting of a list of lemmatized words, and spelling variation rules. It is exactly what we had to produce for IMPACT. So adapting the lemmatizer means producing the IMPACT data, or conversely, producing the IMPACT data means adapting LGeRM to late Renaissance French. The main difference is that LGeRM doesn't use frequency of words, and does use a theoretical lexicon (not necessarily attested in a corpus).

3.6 German

3.6.1 Period and type of material

Our corpus consists of 510 texts varying in length and including different genres. It contains 3,552,690 tokens (words in running text) and 369,730 types (unique words) in total. As the texts originate from 1350-1950, our corpus contains material both from the Early New High German period (1350-1650) and the New High German period (since 1650), covering all subperiods as well.

3.6.2 Peculiarities of language and spelling

As with other languages, we differentiate between two kinds of historical variants: in many (but by far not all) cases, a historical word form w_{hist} can be derived from a corresponding modern full form w_{mod} in a regular way by applying simple rewriting *patterns*. For example, the pattern "*t_th*" explains the difference between the modern word form "Turm" and the historical spelling "Thurm" (English: "tower"). We collected a set of 140 of such manually confirmed rewrite rules. These rewrite rules explain quite a large amount of historical variants in newer texts, the "regular" part of historical spelling variation. In other cases, especially when looking at very old texts, the pairs of modern and historical forms are not connected in a regular, pattern-based way. These historical word forms are called "irregular".

3.6.3 Typefaces, scripts, special glyphs

Modern German has four "special" characters: ä (Ä), ü (Ü), ö (Ö) and ß. In historical texts, we found a large amount of other characters that are not in use anymore. Most of them are combinations of common characters with special diacritics, for example Ůͤ, ā, aͤ, å, ē, eͤ, e̊, ı̄, ỉ, m̄, mͤ, n̄, ō, oͤ, t́, t̉, ū, uͤ, ů, w̃, ȳ.

Especially when going back in time, historical German texts are printed in a Gothic font type, which causes problems for the OCR. Furthermore, the Gothic typeset contains outdated characters. For example, the usage of the "long s" (ſ) was very common until around 1900, and, as in many other languages, we find different obsolete ligatures. In older texts, punctuation also differs from its current use: "virgels" (/) are used to indicate any kind of unit of meaning. In early prints, not only the hyphen "-", but also the symbol "=" marked line breaks. On the other hand, in many cases no hyphenation marks were used at all.

3.6.4 Resource situation

The contemporary language lexicon we use is the CISLEX (Guenthner 1996, Maier-Meyer 1995), a large morphological lexicon for modern German with over two million entries. Each entry represents a word form with associated information. The information attached to a full form includes the underlying lemma, part-of-speech (POS) category, and full morphological analysis.

Our ground truth corpus contains 510 texts from 1350-1950, with currently 3,552,690 tokens and 369,730 types. When creating the corpus, we first searched for freely available ground-truth quality material. The first version of our corpus consisted of:

- The *Bonner Frühneuhochdeutschkorpus:* 40 sources from 1350 to 1700, sorted by language regions.[25]
- The *GerManC Corpus:* 50 newspaper texts of five regions, dating from 1650 to 1800.[26]
- Texts from the *Wikisource Project*: a manually selected sample of 53 German texts, proofread twice, from 1504 to 1904.[27]
- The *Historisches Korpus IDS:* The Institute for German Language (IDS Mannheim) kindly supported our work by providing their corpus for lexicon building.[28]: re-keyed texts of various lengths ranging 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.
- Especially in the sixteenth century, the resulting collection was quite sparse. To close the gaps, we compiled a selection of documents for Early New High German in collaboration with the Bavarian State Library (Bayrische Staatsbibliothek ¨C BSB). The documents were randomly selected from digitized images of the sixteenth and seventeenth-century collection of the BSB and were keyed by a service provider who had access to the electronic images of the scanned books. Our collection was thus expanded by:

[25] http://www.korpora.org/Fnhd/

[26] (Scheible et al. 2011)

[27] http://www.wikisource.org

[28] http://www.ids-mannheim.de/ll/HistorischesKorpus

- The *BSB-LMU corpus:* 101 works with 1,766 pages, adding up to approximately 858,000 tokens. The material was randomly selected from digitized images of the sixteenth and seventeenth-century collection of the BSB. The texts are all derived from a wider focus area connected to theology. Latin documents have been excluded. The documents were manually re-keyed by service providers who had access to the electronic images of the scanned books (cf. above). The keyed material makes up the largest electronic corpus available for research on Early New High German so far.

For all included texts, we store the date of origin and, if available, other metadata such as author, place of origin and genre.

3.7 Polish

Polish is a West Slavic language with about 38 million native speakers in Poland and some users abroad. Like all Slavic languages, it has a rich inflexion, which is now well described formally but only in its contemporary form. For written texts, the Latin alphabet has been used from the very beginning, i.e. the twelfth century, although extended and adapted to Polish phonology. As the phonology changed over the centuries, the present orthographic system is primarily a mixture of phonetic and morphological principles, but some often-used words preserved historical spelling; there are also some orthographic rules which are purely conventional (with conventions changing perhaps too often).

3.7.1 Period and type of material

The ground truth material for Polish consists of books published from 1617 to 1756, the Digital Library of Polish, and Poland-Related News Pamphlets from 1570 to 1728.

3.7.2 Peculiarities of language and spelling

The contemporary Polish alphabet is composed of 32 letters, the 23 standard Latin ones (without q, v and x, which are used only in foreign words) and additional letters being described by the following Unicode names:

Ą	LATIN CAPITAL LETTER A WITH OGONEK
Ą	LATIN SMALL LETTER A WITH OGONEK
Ć	LATIN CAPITAL LETTER C WITH ACUTE
Ć	LATIN SMALL LETTER C WITH ACUTE
Ę	LATIN CAPITAL LETTER E WITH OGONEK
Ę	LATIN SMALL LETTER E WITH OGONEK
Ł	LATIN CAPITAL LETTER L WITH STROKE
Ł	LATIN SMALL LETTER L WITH STROKE
Ń	LATIN CAPITAL LETTER N WITH ACUTE
Ń	LATIN SMALL LETTER N WITH ACUTE
Ó	LATIN CAPITAL LETTER O WITH ACUTE
Ó	LATIN SMALL LETTER O WITH ACUTE
Ś	LATIN CAPITAL LETTER S WITH ACUTE
Ś	LATIN SMALL LETTER S WITH ACUTE
Ź	LATIN CAPITAL LETTER Z WITH ACUTE
Ź	LATIN SMALL LETTER Z WITH ACUTE
Ż	LATIN CAPITAL LETTER Z WITH DOT ABOVE
Ż	LATIN SMALL LETTER Z WITH DOT ABOVE

Actually all those letters are still not sufficient for Polish phonology, so additionally same digraphs are used: *cz, ch, dz, dź, dż, sz, rz*. A special category of digraphs consists of a consonant and the letter i, the latter being both a vowel and a palatalization mark.

The most important difference between historical and contemporary language is the consequence of the phonology changes. As Polish had more vowels than it has now, the following letters were needed:

Á	LATIN CAPITAL LETTER A WITH ACUTE
á	LATIN SMALL LETTER A WITH ACUTE
É	LATIN CAPITAL LETTER E WITH ACUTE
é	LATIN SMALL LETTER E WITH ACUTE

Due to more subtle phonology changes, the letters for some palatalized consonants, e.g.

Ẃ	LATIN CAPITAL LETTER W WITH ACUTE
ẃ	LATIN SMALL LETTER W WITH ACUTE

are no longer used.
On the other hand, in historical texts some contemporary letters are absent and therefore represented by other, consequently ambiguous, letters: historic i may represent contemporary i or j; similarly, historic y may represent y or j.

There are also some quite trivial differences, like the use of long s.
The description given above is not exhaustive, there are also differences in word boundaries, inflectional endings etc.

3.7.3 Typefaces, scripts, special glyphs

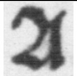
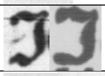
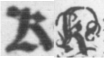
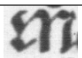
At first the primary script for Polish texts was Gothic (black letter), while Roman type was used, according to its name, for Latin quotations. Later Gothic went out of use and Roman type became dominant.

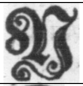
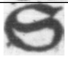
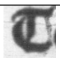
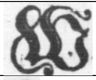
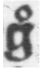
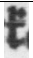
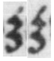
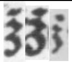
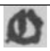
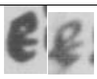
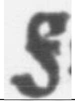
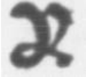
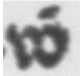
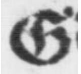
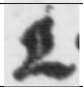
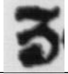
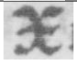
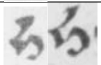
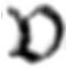
There is a lot of variability and inconsequence in the printed texts, which may be due to the restrictions caused by the repertoire and the size of the fonts used (many printers were immigrants and, at least inititally, used the fonts brought from their country of origin). Not only is *vv* sometimes used for *w*, but the acute accents may have the shape of the macron, etc.

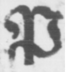
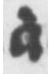
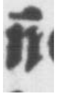
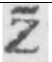
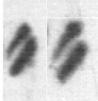
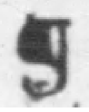
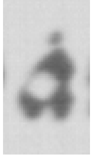
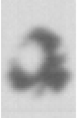
The most prominent special glyph needed for Polish is the ligature of long s and small ł.

Three lists of glyphs have been kindly provided by Mr. Tomasz Parkoła from the Poznań Supercomputing and Networking Center.

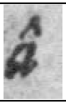
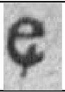
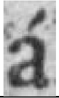
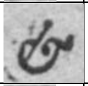
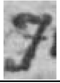
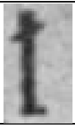
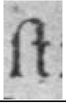
Black letter glyphs:

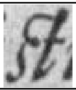
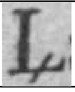
No.	*Glyph*	*Character name*	*Characte r*	*Character code (Unicode)*	*Base character*	*Combining character*
1		Latin capital letter a	A	0041	A	
2		Latin capital letter I	I	0049	I	
3		Latin capital letter k	K	004B	K	
4		Latin capital	M	004D	M	

		letter m				
5		Latin capital letter n	N	004E	N	
6		Latin capital letter s	S	0053	S	
7		Latin capital letter t	T	0054	T	
8		Latin capital letter w	W	0057	W	
9		Latin small letter g with ring above	g̊	???	g	0366
10		Latin small letter k	k	006B	K	
11		Latin small letter z	z	007A	z	
12		Latin small letter z with acute	ź	017A	z	0301
13		Latin small letter z with dot above	ż	017C	z	0307
14		Latin capital letter o	O	004F	O	
15		Latin small letter e with stroke	ɇ	0247		
16		Latin capital letter f	F	0046	F	
17		Latin capital letter y	Y	0059	Y	
18		Latin small letter c with dot above	ċ	010B	c	0307
19		Latin small letter w with acute above	ẃ	1E83	w	0301
20		Latin capital letter g	G	0047	G	
21		Latin capital letter l	L	004C	L	
22		Latin capital letter z	Z	005A	Z	
23		Latin capital letter x	X	0058	X	
24		Combining latin small letter o	ͦ	0366	none	
25		Latin capital letter h	H	0048	H	
26		Latin capital letter u	U	0055	U	

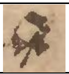
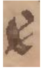
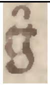
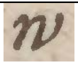
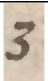
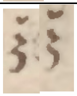
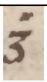
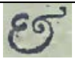
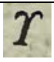
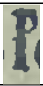
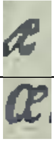
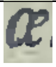
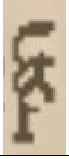
27		Latin capital letter p	P	0050	P	
28		Latin small letter a with grave	à	00E0	a	0300
29		Latin small letter n with macron	n̄	???	n	0304
30		Latin small letter z with tilde	z̃	F516 (in Alethia PUA)	z	0303
31		Double oblique hyphen	⸗	2E17	⸗	
32		CURVED STEM PARAGRAPH SIGN ORNAMENT	❡	2761	❡	
33		Latin small letter a with dot above	ȧ	0227	a	0307
34		Latin small letter a with stroke	ⱥ	2C65		

Glyphs from the "Nowe Ateny" encyclopedia

No.	Glyph	Character name	Character	Character code (unicode)	Base character	Combining character
1		Latin small letter a with circumflex	â	00E2	a	0302
2		Latin small letter e with stroke	ⱥ	2C65		
3		Latin small letter a with acute	Á	00E1	a	0301
4		Ampersand	&	0026		
5		Latin capital letter J	J	004A		
6		Latin small letter l with stroke	ł	0142		
7		Latin small ligature long s t	ﬅ	FB05		

8		Latin small ligature s t	st	FB06		
9		Latin capital letter L with stroke	Ł	0141		

Glyphs from other books in the Polish Ground Truth set.

No.	Glyph	Character name	Character	Character code (unicode)	Base character	Combining character
1		Latin small letter a with ???	ⱥ	2C65		
2		Latin small letter e with ???	ɇ	0247		
3		Latin small letter g with inverted breve	g̑	?	g	0311
4		Latin small letter u with macron	ū	016B	u	0304
5		Latin small letter w	w	0077	w	
6		Latin small letter z	z	007A	z	
7		Latin small letter z with caron above	ž	017E	z	02C7
8		Latin small letter z with dot above	ż	017C	z	02D9
9		ampersand	&	0026		
10		Latin capital letter y	Y	0059		
11		Latin small letter l with stroke	ł	0142		
12		Ligature ae	æ	00E6		
13		Ligature oe	œ	0153		
14		Exclamation mark	!	0021		
15		Latin small letter s followed by Latin small letter z	sz			

3.7.4 Resource situation

There are practically no historical corpora for Polish, which poses a problem for historical linguistics and language technology. This means we had to rely on dictionaries and (modern) lexica.

The following resources have been used directly or indirectly to build the Polish IMPACT lexica:

- *Dictionary of 17th and early 18th-century Polish* (Słownik języka polskiego XVII i 1. połowy XVIII wieku)
- *Linde's dictionary* (1807-1814)[29], in particular the *index a tergo* published in 1965
- *Doroszewski's dictionary of Polish* (Słownik języka polskiego pod red. W. Doroszewskiego), in particular the *index a tergo* digitalized by Robert Wołosz and the Schematic *index a tergo* of Polish word forms digitized by Zygmunt Saloni and Krzysztof Szafran.
- *Grammatical dictionary of Polish* (Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, Robert Wołosz: Słownik gramatyczny języka polskiego, Wiedza Powszechna 2007 , book with CD), in May 2011 the linguistic data were released on 2-clause BSD license: 5 086 141 inflectional forms with grammatical information: http://sgjp.pl/

The following resources for historical and modern Polish are of interest, but have not been used for various reasons:

- *Dictionary of Old Polish*: (Słownik Staropolski, from first written texts to the end of 15th century) ISBN 978-83-04-00472-6
- *Dictionary of sixteenth-century Polish* (Słownik polszczyzny XVI wieku[30])
- *Great dictionary of Polish* (wielki słownik języka polskiego); Work in progress (2007-2012). Primary form: an Internet website[31].

3.8 Slovene

Slovene is a South West Slavic language, with about 2 million speakers. Like other Slavic languages, it has rich inflection, i.e. words belong to paradigms with a complex system of stems and endings, which is one of the first barriers to processing (lemmatizing, tagging) the language. These levels of computational processing had been investigated for the contemporary standard language but, prior to IMPACT, not for historical language.

3.8.1 Period and type of material

Apart from about 40 pages from a sixteenth-century and a seventeenth-century book, the dataset for historical Slovene contains material published from the second half of the eighteenth century to the end of the nineteenth century. The material consists mostly of books, some written in Slovene, representing the more influential books of the time (mostly religious works, but also a play and a cookbook), others translated into Slovene from German. The dataset also includes the publication of one daily newspaper, containing pages from selected years between its founding year 1843 and 1857.

3.8.2 Peculiarities of language and spelling

Today, Slovene is written in the Latin alphabet, containing 25 letters, including č, š, ž, which are the only three not included in the ASCII character set. This is the so-called Gaj's alphabet, which was introduced in the middle of the nineteenth century. Before that, the so-called Bohorič alphabet was used, which was modelled on German. In the mid-nineteenth century two other alphabets were also briefly used, but were soon discontinued, and the Slovene dataset has no examples. The following table shows the mapping of Bohorič alphabet letters to the letters of contemporary Slovene:

Majuscule	Miniscule	Modern Slovene
Z	z	c

[29] Słownik języka polskiego przez M. Samuela Bogumiła Linde, Warsaw, 107-1814, digitized version at for instance http://kpbc.umk.pl/publication/8173

[30] http://www.spxvi.edu.pl/spxvi/

[31] http://www.wsjp.pl/

Zh	zh	č
S, ˛S	ſ	s
Sh, ˛Sh	ſh	š
S	s	z
Sh	sh	ž

It should be noted that the mapping is relatively complex, both from a technical and a textual-comprehension point of view. As can be seen, the same letter combinations (in particular “Sh”) can be mapped to more than one modern equivalent, and the same letters (z, s) now indicate different sounds, making the texts written in Bohorič alphabet difficult to read for today’s readers.

As with other languages, the differences between historical and contemporary language fall into several classes:

- Spelling change:when, apart from the change of alphabet/changes in the alphabet, variable and changing spellings were used either idiosyncratically for specific words, or sounds were written differently from today, or the sounds themselves have changed in the meantime, as well as their spelling. Word boundaries were also often different from contemporary Slovene, especially in function words. Generally speaking, today’s function words / morphemes are more fusional, (“nase”, rather than “na se”), although changes in the other direction can also be observed (“ne bo”, rather than “nebo”).
- Morphological change: when either the morphosyntactic properties of the word have changed (e.g. a masculine noun becomes feminine) or the inflectional endings have changed (“-i” rather than “-u”).
- Lexical change: when words become extinct as others take their place, or when they show semantic shift.

3.8.3 Typefaces, scripts, special glyphs

In the Slovene data there are no special issues with the typefaces; black letter and gothic are not used, and, apart from the factors due to the age and printing process, the typefaces are close to modern ones.

Modern Slovene has three “special” characters, in particular č, š, and ž (and Č, Š, Ž), while the Bohorič alphabet had, like many others, the long s, which is written as “ſ” in lower case, while its upper-case equivalent in Slovene was written as a combination of a symbol resembling an inverted comma followed by capital S. In IMPACT GTD as well as other historical text data, this is encoded as the Unicode Character 'OGONEK' (U+02DB), i.e. “˛S”. This is arguably incorrect, although we have not managed to find a better character yet. As an interesting aside, quite a few capital S following the ogonek were in the GTD found to be actually encoded as the Unicode Character 'CYRILLIC CAPITAL LETTER DZE’ (U+0406), which has the same glyph as the Unicode Character 'LATIN CAPITAL LETTER S’.

Slovene has no di-graphs, although some ligatures were used in printing. These were, in the production of GTD, also encoded as ligature characters. However, when converting the GTD data into the corpus format, they were decomposed into constituent characters. In particular, the following ligature characters are used in GTD, mostly coming from MUFI and encoded in the Unicode Private Use Area in Aletheia:

- Latin small ligature long s + i
- Latin small ligature long s + long s
- Latin small ligature f + f
- Latin small ligature f + i

3.8.4 Resource situation

3.8.4.1 Proof-read historical texts

The available proof-read texts for historical Slovene come from several sources:

1. AHLib digital library, available prior to the involvement of JSI in IMPACT and produced in cooperation with the Austrian Academy of Sciences and JSI. The proof-read part of this DL consists of 90 complete translated books (2,5 million tokens) from 1847-1918. The books are encoded in TEI P5 and richly structured, and have associated facsimiles.
2. IMPACT NUK GTD, the IMPACT ground truth data produced by the National and University Library (NUK), containing complete books mostly from 1750-1850, previously uncorrected books from AHLib, and selected issues from one newspaper. The complete GTD comprises just under 5,000 pages. The data is encoded as Page XML with associated facsimiles. It should be noted that this format is oriented towards OCR technology rather than language explorations.
3. Google ZRC dataset consists of proof-read texts produced by the Scientific Research Center of the Slovenian Academy of Sciences and Arts (ZRC), and funded by the Google award “Developing historical models of Slovene”. The dataset contains either complete or sampled, mostly older books (sixteenth, seventeenth century). Currently, only sampled pages from 3 older religious books are available, encoded in TEI P5.
4. The WikiSource transcribed books, an ongoing project on Slovene literary classics at the Ljubljana University where raw OCR of books and other materials is being hand-corrected by students. Currently it contains about 500 publications (8 million tokens). This corpus contains, in general, more recent texts than AHLib, most of them from the late nineteenth and early twentieth century. We have downloaded the currently finished transcriptions and turned them into a uniformly encoded TEI P5 corpus.

3.8.4.2 Contemporary language lexicon

In order to both process historical lexica and compare them to the contemporary standard, a lexicon of contemporary language is needed. For Slovene we have at our disposal the lexicon extracted from the FidaPLUS reference corpus (600 million tokens), which contains texts from 1991 onwards. This is a large lexicon with over 3 million entries of the form

Word form / Lemma / Morphosyntactic Description / Corpus frequency.

For historical texts, the main problem with this lexicon is that it is too large, i.e. it contains many false friends for historical words. For example, “sim” is a historical form of the contemporary “sem” (I am), but “sim” exists in the lexicon as the SIM card of a mobile phone – thus the word is incorrectly classified as a modern noun (rather than an archaic form of the auxiliary verb).

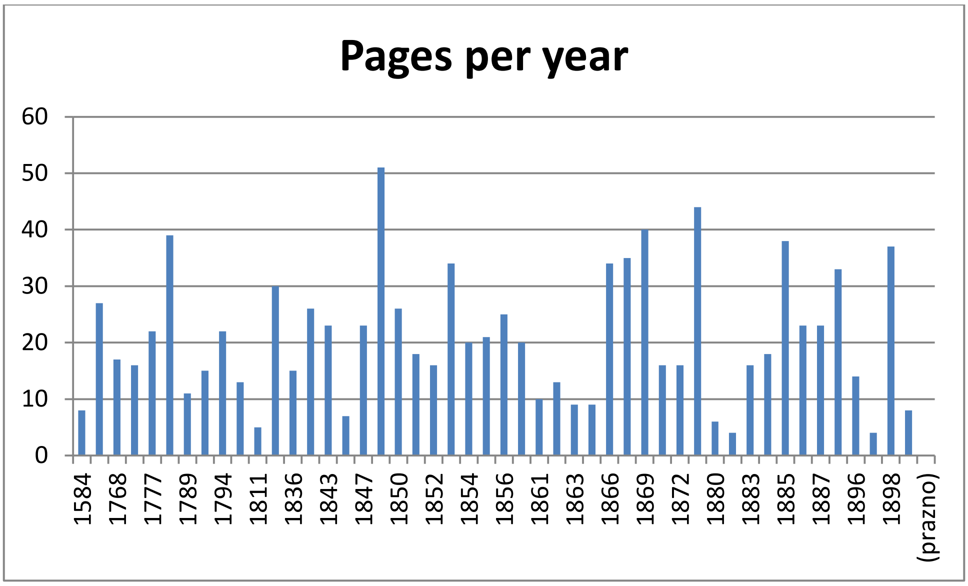
3.8.4.3 Annotation tool

Prior to IMPACT, JSI had developed the annotation tool ToTaLe, which performed tokenization, tagging and lemmatization on texts of contemporary Slovene. For IMPACT, this tool was extended with an extra step. After tokenization, first of all the spelling of the word forms is modernized.Only then are they passed on to the tagging and lemmatization modules. The basis of the tool is the LMU Vaam (Variant approximate matching) library, which is able to transcribe archaic forms into contemporary ones, provided with a list of transcription patterns and a contemporary language lexicon as background resources. This tool, ToTrTaLe, is being updated in the scope of the IMPACT project, with regard to its functionality as well as to its background resources (Erjavec, 2011). For further information, please refer to section 5.7.

3.8.4.4 The reference corpus of historical Slovene

At the mid-point of the Slovene involvement in the IMPACT project, JSI, together with ZRC, obtained the Google Award for developing models of historical Slovene. This provided extra funds for work on the language resources, and in particular, it made it possible to create a reference corpus of historical Slovene, containing 1,000 pages (250,000 tokens) of material sampled from the AHLib digital library and the available IMPACT NUK GTD and Google ZRC dataset.

The pages were specifically selected to cover as many interesting publications and time periods as possible (see the table below).



For the sake of convenience, the corpus has been divided into four parts, roughly corresponding to the sources and time periods:

- goo168: consists of selected consecutive pages of the ZRC books and contains the oldest material from the corpus, i.e. samples from three religious books from the end of the sixteenth, seventeenth and eighteenth centuries, 74 pages, and 20,000 word tokens, in all.
- goo18B: sampled pages from books from the second half of the eighteenth century, obtained from the NUK GTD, 116 pages, and 16,000 word tokens, in all.
- goo19A: sampled pages from books and newspaper volumes from the first half of the nineteenth century, obtained from the NUK GTD, 180 pages, and 55,000 word tokens, in all.
- goo19B: sampled pages from books and newspaper volumes from the second half of the nineteenth century, obtained from the NUK GTD and the AHLib corpus, 630 pages, and 150,000 word tokens, in all.

For further discussion of the annotation of this corpus, please refer to section 5.7.

3.9 Spanish

Spanish is an Indoeuropean language belonging to the Romance group. It has about 400 milion native speakers and 50 - 100 million of non native. It is the second native language behind Mandarin Chinese and the second one more studied behind English, with about 20 million of students. In Spain is spoken as native language by 42 million of people. Spanish is official language in 21 countries apart of United Nations, European Union, Organization of American States, and other organizations.

Spanish language is lexically more similar to Italian and Portuguese than to other Romance languages as French or Romanian.

3.9.1 Period and type of material covered

Fourteen works of Spanish Literature and a dictionary (consisting of 6 volumes) were selected for the IMPACT Demonstrator dataset. Most books are from the sixteenth or seventeenth century, the

Spanish Golden Age. They are mostly literary works: religious, plays, novels, poetry... Just one belongs to eighteenth century, as does the *Diccionario de Autoridades*.

Two books are from America: *Cartha Athenagorica* by Sor Juana Inés de la Cruz and *Comentarios reales* by Inca Garcilaso de la Vega. They were selected in order to register the vocabulary of Spanish in Latin America.

3.9.2 Peculiarities of Spanish language and spelling in the period covered

Nowadays, Spanish is written using a variation of the Latin alphabet which is composed of 27 letters, including the character *ñ* (eñe). From the nineteenth century onwards Spanish alphabet has contained 29 letters, the current digraphs *ch* and *ll* are considered as one letter.

Spanish language standardization started in the eighteenth century. From 1854 onwards, Spanish orthography has been the responsibility of the Royal Academy. Since then the orthography has undergone limited changes, most of them due to phonological criteria.

Spanish orthography has since the sixteenth century undergone the following main changes:

- In Old Spanish there were two palatal sounds, voiceless, written as *x* and voiced represented by *j*. In the sixteenth century the voiced sound disappears and both are written as *j*. In the IMPACT dataset for Spanish, which includes works of the sixteenth and seventeenth century, we still find some words written in the Old Spanish way, as in *debaxo* instead of *debajo*.
- The letter *ç* is not used in modern Spanish, but *z,* as in *coraçon* - *corazón*.
- *ze*, *zi* are now written as *ce*, *ci,* as in *dezir* - *decir*.
- Words written with *qua* are now written with *cua,* as in *qual* - *cual*
- Double consonants are now simplified, as in *grammatica* - *gramática*
- Greek-Latin digraphs are reduced: *ch* - *c (christiano - cristiano), ph* - *f (philosophia - filosofía), th - t (theatro - teatro).*
- Numbers higher than 10 and 20 are now written as one word, while they were written as three words in Old Spanish: *diez y siete* is now *diecisiete* and *veinte y seis* is *veintiséis*.
- Some latin consonant groups are still present in older Spanish and have currently disappeared, as in *redemptor* - *redentor* or *subjeto* - *sujeto*.
- In Old Spanish very often *h* was used to indicate diaeresis instead of acute, as in *ohir* - *oír*.
- In Old Spanish the use of *v* instead of *u* as in *vno* - *uno* and the use of *u* instead of *v* as in *nueuo* – *nuevo* occurs quite frequently.
- The use of diacritics in general is Old Spanish quite different from Modern Spanish. Currently only the acute is used. In the IMPACT dataset for Spanish however, we find examples of circumflex and grave and the position of the accent is also different, as in *segúro* - *seguro, abalançò* - *abalanzó.*
- The difference between Modern Spanish and older Spanish is also visible for verbs, for example:
  - Drop of final *d* in imperative: *volvé* - *volved*
  - Use of *u* or *v* for the imperfect indicative instead of current *b*: *cantaua* - *cantaba* or *ocupavan* - *ocupaban*.
  - Use of the archaic imperfect subjunctive: *fuerades* - *fuerais*.

Apart from changes in orthography, there is also lexical, semantic and morphological variation, but not significant enough to be mentioned here.[32]

[32] See also : http://en.wikipedia.org/wiki/Spanish_language
http://en.wikipedia.org/wiki/History_of_Spanish;
http://en.wikipedia.org/wiki/Old_Spanish_language
http://en.wikipedia.org/wiki/Spanish_orthography

3.9.3 Typefaces, scripts, special glyphs

For Spanish we did not have issues with typefaces, since we neither had gothic nor blackletter font in our data.

Modern Spanish has just one special character, *ñ*, but in the period covered, as we have seen in the previous section, there are more:

- *ç* [ʃ] replaced nowadays by *z* or *c* [θ].

- Tildes: We have some different uses for the tildes:

    - Over a n (ñ): It is used in order to represent two sounds, a double *n* [n] and the current *ñ* [ɲ]
    - Over a *q* (*q̃*): used as abbreviation of *que* as a word or as a syllable as in *q̃dar - quedar*.
    - Over a vowel as an abbreviation, it indicates that after should be a nasal consonant as in *siẽpre - siempre*, *ultimamẽte - últimamente.*
    - Over any letter: used, specially over *r*, to represent an abbreviation as in *vr̃a - vuestra* or *mr̃d - merced.*

- Long *s* (ſ), replaced in Modern Spanish by small *s*.

- Ligatures:

    - Latin small ligature c + t: *efecto*
    - Latin small ligature c + h: *noche*
    - Latin small ligature f + f: *offender - ofender*
    - Latin small ligature f + i: *fingia*
    - Latin small ligature f + l: *flores*
    - Latin small ligature i + j: *iij*
    - Latin small ligature l + l: *cauallo - caballo*
    - Latin small ligature long s + s: *paſsè - pasé*
    - Latin small ligature long s + t: *paſtor - pastor*
    - Latin small ligature long s + i: *aluſion - alusión*
    - Latin small ligature long s + long s: *aſſegurar - asegurar*
    - Latin small ligature long s + long s + i: *grandiſſimo - grandísimo*
    - Latin small ligature s + t: *vueﬆro - vuestro*

- *&c* meaning the abbreviation *etc*. (*etcétera*).

- Other non alphabetic special characters are:

    - *Punctus elevatus* (⸵): This is a special character not used in Modern Spanish
    - Inverted question and exclamations marks (¿¡): They are still used.

3.9.4 Resource situation: corpora, lexica

3.9.4.1 Proof-read historical texts

The available proof-read texts for historical Spanish have come from two sources:

- A collection of 86 complete books (almost 2 million tokens and 90.000 wordforms) from late 15th Century - 17th Century, in TEI and coming from the Biblioteca Virtual Miguel de

Cervantes (BVMC) - available prior to the involvement of the University of Alicante (UA) in IMPACT.

- IMPACT ground truth material produced by the UA from images provided by the Biblioteca Nacional de España (BNE), containing 14 complete books mostly from sixteenth and seventeenth century, belonging to the *Siglo de Oro Español* and a dictionary, *Diccionario de Autoridades,* published in 6 volumes in eighteenth century. The GT dataset consists of ca. 11.000 pages containing about 6 million tokens.

3.9.4.2 Contemporary language lexicon

For Spanish we have used the modern lexicon from Apertium. Apertium is a freely available open source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan), developed by the Transducens research group at the *Departament de Llenguatges i Sistemes Informàtics* of the University of Alicante in collaboration with Prompsit Language Engineering. The *Apertium* lexicon contains over 700.000 entries of the form *word-form / lemma / simplified PoS.*

3.9.4.3 Annotation tool

In order to annotate the proof-read historical texts we used the IMPACT tokenizer and the Lexicon Tool (CoBaLT), both developed by INL.

3.9.4.4 The reference corpus of historical Spanish

The corpus is for convenience divided in several parts corresponding to its origin:

- BVMC Corpus: Consists of texts provided by BVMC almost 2 million tokens and 90.000 wordforms.
- GT Development Corpus: Consists of pages from GTD of Development subset, about 600.000 tokens and 60.000 types.
- Dictionary Development Corpus: Consists of pages from *Diccionario de Autoridades* GTD; although it is part of the development subset, due to its size we have annotated it separately. It consists of about 4 million tokens and 300.000 wordforms.
- IR Evaluation Corpus: Selected pages from Evaluation dataset. It is a completely annotated corpus of 10.000 tokens and about 4.000 types.

4 Building the lexica and the spelling variation models

4.1 Bulgarian

4.1.1 Lexicon building with LeXtractor

The Bulgarian involvement in IMPACT was marked by the lack of IMPACT GTD data until the end of September 2011 and the use of the LeXtractor tool from LMU for lexicon building. A corpus of 7 different sources covering the end of the nineteenth century was compiled and OCR-ed with Abbyy's FineReader 9.0 using specific training patterns for obsolete characters (e.g. ѭ, ѫ, ѣ). The most frequent words were manually checked and the resulting corpus of 1765 pages (1,529,851 words) was imported into LeXtractor. As a background contemporary Bulgarian language lexicon we used the BAS Dictionary. The transcription rule-set was based on 8 expert-knowledge rules and about 30 empirically established rules, by observing the effect of each proposed rule on the recognition of historical words in the corpus.

In the end of June 2011 the corpus was extended and it currently amounts to 2676 pages (about 2,221,000 words). In this way we managed to increase the frequency of correctly recognized words while suppressing the erroneous ones. Another approach we took in order to deal with OCR errors was to establish simple but safe rules which identify the errors automatically.

In this way we compiled a list of erroneous words that were automatically excluded from the system by the team at LMU. When the GT data arrived, we tested for correctly recognized, but erroneously neglected, words. We found about 300 of such cases, which were reactivated in the system so that they will be processed.

On this basis we proceeded with lexicon building, concentrating on the high-frequency words in the OCR corpus, either those that were recognized as variants of modern words or those that were not recognized by the system. In this way a lexicon of about 26,000 words was developed.

4.1.2 Lexicon of historical words

The current lexicon consists of 28,857 lexical entries developed in LeXtractor and the lexicon that can be automatically extracted from the manually validated tokens from the reference corpus. At the time of writing, the size of the historical lexicon extracted from the manually validated corpus tokens is given in the following table:

	Historical Lexicon
Lex. entries	28,857
Word forms	26,148
Normalised	25,861
Modernized	21,115
Lemmata	11,090

The first row gives the number of lexical entries, the second the exact number of word forms as they appear in the corpus, the third the number of normalized word forms, i.e. lower-cased, the fourth the number of modernized word forms, and the fifth the number of modern lemmata.

The lexicon is currently available as LeXtractor and TEI P5 XML and in the IMPACT database structure; in addition to the information in the table, it also contains the number of times a particular lexical item occurs in the corpus and the number of times it has been validated by hand, as well as the listing of all the corpus elements (page ids) in which the particular item has been attested. As these identifiers also contain the year of publication for each element, it is then easy to provide an estimated time period in which a particular lexical entry was used.

4.1.3 Spelling variation models

As mentioned above, the initial spelling variation model for LMU was based on experienced knowledge of several certain transformations of obsolete characters in nineteenth-century Bulgarian. These rules were subsequently extended empirically, based on the OCR corpus and the BAS Dictionary. The results obtained in this way were manually controlled before being integrated in the system.

4.2 Czech

The core general lexica for Czech are based on the following data:

a) dictionaries of J. Dobrovský (published in 1802 and 1821), J. Jungmann (1835 –1839) and F. Š. Kott (1878-1897), with the headwords of these dictionaries expanded into full paradigms;

b) nineteenth-century texts from the text bank of the Czech National Corpus;

c) nineteenth-century texts texts from the development subset of the Czech National Library Demonstrator set for IMPACT.

The building of lexica based on Jungmann’s and Kott’s dictionaries has been carried out as follows:

1. Proofreading the OCR texts of the two dictionaries (dictionary headwords only)
2. Using Czech National Corpus nineteenth-century texts (200,000 word forms, apart from IMPACT GT set) to add attestation data to the rather too elaborate word lists extracted from the dictionaries (goal: elimination of rare words)
3. Expanding the keywords of the two word lists into full paradigms
4. Testing the coverage of the resulting two lists of word forms on part of GT text set
5. Adjusting the word lists accordingly to reach the target coverage of 80%

4.3 Dutch

The Dutch IR lexicon has been built by means of the IMPACT dictionary attestation tool from the quotations of the WNT (Dictionary of the Dutch language, cf. above). The lexicon currently contains 475,498 distinct word forms, 215,180 lemmata, and 558,438 distinct lemma/word form combinations, with 1,636,709 attestations. After the automatic matching of the headwords in the quotations, the manual processing took about 775 hours.

The OCR lexicon used in this report is corpus-based, using the large historical corpus from the DBNL (Digitale Bibliotheek voor de Nederlandse Letteren, Digital Library for Dutch Literature). An alternative OCR lexicon based on the contents of the IR lexicon is also available.

A reference set of 10,000 pairs (modern word plus historical word), taken from the WNT-based lexicon, has been built for the inventory of historical spelling variation. Based on this, a set of spelling variation rules has been constructed.

4.4 French

The global method consisted in the adaptation of LGeRM for late Renaissance French. The results were exported in accordance with IMPACT tools specifications.

The building of the historical lexicon can be divided into two periods: before GT and after GT. We were a little bit optimistic about our textual database Frantext content. We provided a first lexicon for LGeRM using Frantext. When we compared this to the GT, we discovered that Frantext texts were partially modernized, so that the coverage rate (although quite high and already meeting IMPACT specifications) was lower than we thought.

4.4.1 Before GT

Our first approach was to use the LGeRM lexicon for Middle French and to add modern (out-of-LgeRM) forms from Morphalou to it. This task was easy, as the lemma forms of DMF are equal to modern entries of TLF/Morphalou whenever possible. The Middle French lexicon has 785,000 inflected forms. We produced a new lexicon with 1,121,000 inflected forms. We tested this lexicon on a first set of texts from the seventeenthe century. The coverage of this initial lexicon was around 95% of words. It allows us to adapt the LGeRM lemmatizer and to add new variation rules.

4.4.2 GT is available

It was necessary to test the lexicon on IMPACT GT to evaluate its accuracy. First finding: our textual database had been partially modernized and there was a lack of U/V I/Y and umlaut variation. Second finding: medieval variation is a lot more complicated. The lexicon coverage was rather good but in terms of Information Retrieval, it produced too much noise. Example: “servent” in medieval French is a form of SERVANT, noun, and of SERVIR, verb (to serve). But for late Renaissance French only the second should be in the IR lexicon. It was quite difficult to reduce noise, and after consultation with the INL, we decided to build a new lexicon, starting directly from Morphalou and trying to archaise modern forms. We built a new text corpus including the initial texts from Frantext, GT, and new texts from Frantext, dating from 1700 to 1740. The whole corpus contains 10,684,781 tokens (including punctuation), 132,005 types (including words split by line breaks, words, ground truth errors).

The initial lexicon is Morphalou. Our goal is to extend the lexicon to produce all types from the textual corpus.

The algorithm to produce a hypothetical lexicon consists of 4 steps. The first step is to apply all existing variation rules to enrich the lexicon. These rules are applied a second time and a third time on the result files produced at each step. The fourth step is to look for all unknown types in the LGeRM knowledge base. The lexicon produced is used to lemmatize the set of types from the corpus. All unknown variant forms are analyzed to find new variation rules or to detect new lemmata (and their inflections).

The process is iterated as often as necessary to cover the whole set of types.

As a result of that operation, we can produce a list of words with lemmata and a list of spelling variation rules. The hypothetical lexicon contains 2,221,000 entries. The coverage of the hypothetical lexicon is around 99,5%. For IMPACT we only use forms attested in corpora.

This method guarantees a/leads to a clean ground truth. All remaining words without lemmata are considered GT errors. So by using LGeRM we can quite easily produce IR Evaluation Sets.

4.4.3 New release

Because of the time constraints in IMPACT, some aspects could not be completed for the first release of the lexicon.

In the first release, we were unable to completely distinguish compound words from words containing a hyphen. In French, the same character is used in both situations, and many compounds are not included in Morphalou. So the frequencies produced in the OCR lexicon need to be corrected (the actual frequencies are not so bad, but could be refined).

Secondly we did not manage to establish a checked list of available lemmata. As a consequence, it cannot be excluded that this release accidentally contained invalid lemmata or invalid parts of speech for a lemma.

Finally, the method proposed by IMPACT excluded two thirds of the GT corpora (the evaluation and demonstration subsets). They contain variant forms and new lemmata that can be included in the lexica.

4.4.4 GT errors

Ground truth was supposed to be clean. But in fact using LGeRM we detected errors. As we were running out of time, we decided not to report them. All corrections were encoded in the TEI XML version of text. Information from the XML Page format was kept (page id, paragraph id), so it would be possible to generate a list of corrections to apply.

We also detected printing errors. Although these are in themselves not a point of interest to IMPACT, for linguistics studies, it is quite necessary to have that kind of information.

4.5 German

To build the IR lexicon of historical German, we developed the LeXtractor-tool. This tool enables the efficient corpus-based construction of historical lexica. It supports web-based collaborative work, and special language technology for constructing entries is integrated to reduce the manual effort of the lexicographers wherever possible. The idea is that whenever possible the tool creates suggestions for lexical encoding of entries and the user just confirms or rejects these suggestions. The creation of the lexical entries is carried out in a frequency-ordered way, giving preference to words with many occurrences, and the lexicographers are guided through a series of interactions that ensures that all possible readings for this string are considered and valid readings are stored in the lexicon. In IMPACT, the tool has been used by other language partners to create their historical lexica.

As to German, 22,800 non modern entries with attestations in the available corpus material have been created up to now. The lexicon contains 20,700 different historical strings, which means we found attestations for approximately 1,1 different readings of a string. 36,800 readings in total have been manually marked as feasible, but 14,000 of them could not be verified in the corpus. Of all 36,800 processed readings, 31,700 are pattern-based and 5,100 are “irregular”. These 36,800 readings point to 19,200 lemmata.

4.6 Polish

Prior to IMPACT, there were practically no historical corpora of Polish, which caused various problems from the very beginning. One of them was the lack of standards for representing old Polish texts in Unicode, as several necessary characters and ligatures are not provided, neither by the Unicode proper nor by Medieval Unicode Font Initiative. Together with our library partner, PSNC, we successively developed the instructions for the ground-truth service provider, analyzed the character histograms and word frequency lists of the GT files with the help of the Poliqarp search engine for DjVu [33] immediately after receiving a new batch and, when needed, recommended assigning a Private Use Area code in the ground truth tool used in the IMPACT project, *Aletheia*[34].

At first the idea to use the dictionary of sixteenth-century Polish (which we also refer to as "Early Middle Polish dictionary") was seriously considered and a case study was made of converting sample entries into IMPACT database format, but the idea had to be aborted for several reasons, including legal ones (the copyright holders were not interested in cooperation).

Our primary resource was the Internet dictionary we shall refer to as the "Late Middle Polish dictionary", its official name being "The dictionary of the Polish language of the sixteenth and the first half of the seventeenth century"[35]. The dictionary is still under development; almost 15 thousand entries are in different stages of preparation. Thanks to the head of the editorial team, Prof. Włodzimierz Gruszczyński, the dictionary has been made available for the IMPACT project in the form of a database dump. Unfortunately the dictionary is underfunded and as a consequence there is no written documentation of the database and the part-time administrator was too busy to provide consultations.

One of the most useful parts of the dictionary database content is the set of over 40,000 quotations from over 700 sources; some quotations are in modernized spelling and some in a spelling close to original. Sizewise, the quotations amount to over 3 MB of characters. To gain a better insight into their nature they were first imported in a "quick and dirty" way into Poliqarp corpus management software[36]. Later the quotations were prepared for import into the IMPACT Attestation Tool, which particularly required cleaning some editorial codes and using appropriate tokenization algorithms. Then the IMPACT Dictionary Attestation Tool was used to produce the lexicon, and this task was completed by the end of June 2011.

Other very useful information extracted from the dictionary database is the list of over 50,000 word forms together with their lemmata (in modernized spelling) and some grammatical information, facilitating automatic lemmatization in particular.

The auxiliary resources used included Krzysztof Szafran's morphological analyzer, which is based on the data gathered by late Prof. Tokarski during the work on the dictionary we shall call "Early Modern Polish dictionary" (about 120,000 entries) , and the word form list of the grammatical dictionary of modern Polish (over 500,000 items, including proper names and words of foreign origin).

To get better coverage of historical lexica we decided to also use Linde's dictionary[37], a large and important nineteenth-century dictionary of Polish. We actually used it indirectly by digitazing the

[33] http://dx.doi.org/10.1007/978-3-642-23160-5_1

[34] http://tools.primaresearch.org:8080/tools/primaweb/tool.php; Clausner et. al. 2011.

[35] http://sxvii.pl/.

[36] http://poliqarp.sourceforge.net/

[37] http://eprints.wbl.klf.uw.edu.pl/view/creators/Linde=3ASamuel_Bogumi==0142=3A=3A.html

index a tergo published in 1965[38], which provided a list of almost 80,000 entries and subentries. The index was scanned and OCR was performed with FineReader 10 Desktop. Then a series of scripts was written to use the inherent redundancy to discover OCR errors: entries should fall within the range specified by the running head and should be listed in reverse alphabetic order. Unfortunately, the entries in the index are interspersed with subentries which do not follow the alphabetic order and are to be distinguished by a smaller print, which was not always recognized as such. As a consequence, a substantial amount of human proof-reading has been necessary. The results were cross-checked with Szafran's analyzer in two modes: the strict mode, which in practice means checking against the Early Modern Polish dictionary, and the guesser mode, which is less reliable. The results have also been checked with the modern analyzer based on the grammatical dictionary mentioned above, which makes them quite reliable. It is worth mentioning that a lot of OCR errors were the result of FineReader outputting the contemporary variant of a word.

As for spelling variation, we approached that problem in an iterative way. Some empirical data have been corrected when attesting the dictionary quotation, but we started with simple and obvious rules. They were used to normalize the IR evaluation dataset before applying the lemmatization process to it. Despite of the simplicity of the rules, the automatic lemmatization, using Szafran's analyzer and the Late Middle Polish dictionary, was quite successful, so preparing the data for IR evaluation with the IMPACT Lexicon Tool turned out to be quick and easy. Next we designed a simple format for spelling variations rules (context to the left and right is defined by regular expressions, and in addition every rule has an optional exception list). We have written down more sophisticated rules, including those known from the literature (e.g. Tomasz Lisowski, Grafia druków polskich z 1611 i 1612 roku. Problemy wariantywności i normalizacji. Poznań 2001), and verified them on the GT texts, using the above-mentioned Poliqarp for DjVu in particular.

4.7 Slovene

The development of the Slovene lexicon of historical words can be divided into two periods. The methods, datasets, workflows and indeed the goals have changed considerably between the two.

4.7.1 Lexicon building with LeXtractor

The first year of the Slovene involvement in IMPACT was marked by the lack of IMPACT GTD data and the use of the LeXtractor tool from LMU for lexicon building. The AHLib DL was imported into LeXtractor, and the FidaPLUS lexicon was used as the background contemporary language lexicon. The transcription ruleset was developed empirically, by observing the effect of proposed rules on recognition of historical words in AHLib via the Vaam library.

JSI and LMU closely collaborated in adapting LeXtractor - which had been previously used only for processing German - to the processing of Slovene.

On this basis we proceeded with lexicon building, concentrating on the high-frequency words in the AHLib corpus, either those that were recognized as variants of modern words or those that were not recognized by the system. In this way a lexicon of 3,000 words was developed.

In this time period, the first version of the ToTrTaLe was also developed, which enabled annotation of historical texts in such a way that the text is tokenized and each word is annotated with its modern equivalent, its morphosyntactic description and its modern lemma. The tool supports TEI P5-encoded texts, in input as well as output (Erjavec, 2011).

4.7.2 Corpus annotation with the CoBaLT lexicon building tool

The second year of the project brought with it significant changes, due to several factors:

- JSI obtained from NUK the first batch of GTD, allowing the corpus to be extended with pre-1850 texts, partly distinguishableby their use of the Bohorič alphabet.

[38] http://eprints.wbl.klf.uw.edu.pl/19/

- The INL lexicon building tool became available, in the sense that its functionality was extended to allow the annotation of historical word forms not only with their contemporary lemma, but also with their contemporary word form, a necessary prerequisite for highly inflective languages such as Slovene.
- With the first 3,000 high-frequency entries it was possible to significantly improve the annotation quality with ToTrTaLe, enabling better automatic annotation of historical texts.
- JSI obtained the Google award, which enabled the hiring ofextra annotators, significantly boosting the work force/manpower available for the task of manually correcting language resources of historical Slovene
- As part of the Google award, JSI also obtained manually corrected page samples of three older Slovene books, from the end of the sixteenth, seventeenth and eighteenth century.

On account of these changes, we decided to concentrate on fully annotating a reference corpus of historical Slovene and extracting the lexicon from this corpus. The corpus was first sampled from the available material (cf. section 4.8.4.4), encoded in TEI P5, and automatically annotated with ToTrTaLe.

The next stage was the manual correction of the annotations automatically assigned to the corpus. This involved:

- correcting the remaining transcription errors in the corpus;
- correcting automatic word tokenization, esp. where a group of historical words is one word in contemporary Slovene or vice versa;
- correcting the automatically assigned modern word form of the historical word tokens;
- correcting the automatically assigned modern lemma of the word tokens;
- correcting the automatically assigned coarse-grained morphosyntactic description of the word tokens;
- and for words that do not have a transcription-based modern equivalent, constructing their hypothetical modern word form and lemma, and specifying the closest contemporary synonyms, together with the sources from which this information was obtained.

We hired the annotators, trained them on sandbox corpora, set up a mailing list for questions and prepared the User manual, Annotator Cookbook and, as the project progressed, a FAQ on annotation questions.

The reference corpus was then imported into the INL Lexicon building tool, and manual annotation started for real.

In tandem with annotation we also developed, in close cooperation with the INL, the TEI P5 import/export format and contributed to the improvement of COBALT by suggesting useful user requests and meticulously reporting any bugs.

4.7.3 Lexicon of historical words

The current lexicon consists of the initial 3,000 lexical entries developed in LeXtractor and the lexicon that can be automatically extracted from the manually validated tokens from the reference corpus. At the time of writing, the size of lexica extracted from the manually validated corpus tokens was as follows:

Goo lexicon	All	Historical
Lex. entries	46,999	16,245
Word forms	44,903	15,715
Normalized	39,766	14,249
Modernized	33,037	11,396
Lemmata	17,902	6,789

The first row gives the number of lexical entries, the second the exact number of word forms as they appear in the corpus, the third the number of normalized word forms, i.e. lower-cased and with removed vowel diacritics (which were extensively and inconsistently used in historical Slovene, but are not used in contemporary Slovene), the fourth the number of modernized word forms, and the fifth the number of modern lemmata. The first column gives the sizes for the full lexicon (discounting numerals) while the second column gives numbers only for those entries in which the normalized form differs from the modernized form, i.e. only for historical words. It should be noted, however, that including words that have not changed in the lexicon also has the advantage of identifying which modern day words were actually used in historical texts.

The lexicon is available in two formats. The first is a simple tabular file; apart from the information shown in the table, it also indicates the number of times a particular lexical item occurs in the corpus and the number of times it has been manually validated, and lists all corpus elements (page ids) in which the particular item has been attested. As these identifiers also contain the year of publication for each element, it is easy to give an estimated time period in which a particular lexical entry was used.

The second format is as structured lexical entries encoded in TEI P5, using the dictionary module. The export of this format is directly supported by COBALT, and we also developed a script to convert this XML into HTML for browsing. The TEI provides a stable storage format, and will serve as a resource for ToTrTaLe, while the HTML enables the inspection of lexical items in a lemma-oriented fashion, to discover remaining problems and mistakes.

After the completion of the reference corpus we imported the complete corpus into the INL lexicon tool (i.e. the complete NUK GTD and the complete AHLib DL), and are concentrating on adding to the lexicon as many words as possible, along with their modern equivalents.

4.7.4 Spelling variation models

As mentioned before, the initial spelling variation model for LMU Vaam was empirically obtained, based on the AHLib corpus, the lexicon of contemporary Slovene, and the usage of Vaam with the newly-developed rules. When the NUK GTD was obtained, these rules were - again empirically - extended in order to deal with older Slovene, , and they seem to give relatively good results.

With the completion of the annotated reference corpus, we will be in the position to automatically induce transcription rules directly from the lexicon, using the INL Spelling variation tool. The plan is to develop sets of rules separately for the four parts of the corpus, i.e. goo168, goo18B, goo19A and goo19B (cf. section 4.8.4.4), as the periods are distinct enough for separate rulesets to be useful.

4.8 Spanish

4.8.1.1 Corpus annotation with the INL Lexicon Tool (CoBaLT)

The development of the Spanish Lexicon has started before the production of the GT dataset, thanks to the availability of the corpus of digital texts of the Biblioteca Virtual Miguel de Cervantes (BVMC). These texts were imported into the INL Lexicon Tool and the *Apertium* lexicon was used as the background contemporary language lexicon. Lexicon building started by annotating the high frequency words first. The result was a lexicon of over 30.000 types and 500.000 tokens coming from the BVMC texts.

As soon as most of the GT data was delivered and approved, it was used to extend the lexicon. Here also, we started with the annotation of high frequency words first.

As we said in section 9.4.4, the Spanish GT corpus was split into four sub corpora and we have semi-automatically annotated in this way. We followed the following process:

- Correcting the remaining transcription errors in the corpus.

- Correcting (or adding) the automatically assigned lemma of the tokens and wordforms.
- Adding modern lemma to the tokens.
- Adding modern equivalent to the tokens and wordforms.

From the GT dataset, we have extended the lexicon with about 3.000 types and 44.000 tokens.

Apart from this, the University of Alicante and INL have been collaborating closely in order to adapt the Lexicon Tool and its tokenizer to the Spanish language.

4.8.1.2 Lexicon of historical words

The current lexicon consists of about 35.000 types and 600.000 tokens developed using the INL Lexicon Tool (CoBaLT). Each type has its modern equivalent word form and is linked to a modern lemma. The Spanish lexicon contains ca. 12.400 lemma's.

4.8.1.3 Spelling variation models

We have extracted spelling variation rules by hand while annotating the corpus.

4.9 Lexicon building: summary

	Non-IMPACT background corpus	Development ground truth	"unclean" (OCR) material	Modern full-form lexicon	Historical Dictionary entry list	Quotations from historical dictionary
Bulgarian	-	+	+	+	-	-
Czech	+	+	-	+	+	-
Dutch	+	-	-	+	+	+
English	-[39]	-	-	-	-[40]	+
French	+	+	-	+	-	-
German	+	+	-	+	-	-
Polish	-	+	+	+	+	+
Slovene	+	+	+	+	-	-
Spanish	+	+	+	+	-	-

Table: resources used for lexicon building

Language	OCR lexicon	IR lexicon: lemmata	Word forms	Lemma / word form combinations
Bulgarian	81589	12 436	29 796	32 947
Czech 1801-1809	311362	16052	311362	321,099
Czech 1810-1842	297122	16056	297122	304711
Czech 1843-1849	178783	9406	178783	183079
Czech 1850+	596663	31954	506663	518628
Dutch	501228	229454	500382	585130
English		297743	532671	874311
English 1580-1720	406296			

[39] The *ECCO* corpus has been considered as source for the OCR lexicon, but it has not been included.

[40] Of course, the OED headword form is included in the lexicon. But the entry list has not been used as a separate starting point for lexicon building.

English 1700-1800	220044			
English 1750-1920	574444			
French	15481	27508	115201	141152
German		19200	22800	36800
German 1500-1600	23139			
German 1600-1700	29805			
German 1700-1800	44880			
German 1800-1900	98028			
Polish	167160	9909	24977	26736
Slovene	231033	30470	49020	66870
Spanish	147192	11846	31584	36857

Table: size of lexica

5 Deploying OCR lexica

5.1 Deployment and evaluation of lexica in OCR

5.1.1 OCR evaluation: use of lexica in OCR

An extensive evaluation of the contribution of the IMPACT lexica to text recognition has been conducted. The evaluation was carried out by comparing FineReader Engine version 10 in its optimal internal dictionary and language setting (for English, French, German and Spanish, already available historical dictionaries in the FineReader SDK distribution have been used) with FineReader using the same internal dictionary combined with an external historical dictionary that wasrun through the FineReader external dictionary interface (cf. section 2.3.2). In most cases, the combination seems to give better results than the use of the historical external dictionary alone. A purely scientific comparison between the current internal lexica and the external IMPACT lexica was not feasible, because we lack information on how the dictionary is used internally, and are not able to convert the IMPACT lexica into the internal format. Moreover, from a practical point of view, we are more concerned with enhancing existing functionality than with replacing it.

The results are based on a word-based, case- and punctuation-insensitive alignment of ground truth and OCR. Since the alignment is done region-by-region, complex layouts can still be more or less evaluated. We developed a custom evaluation tool because it enables us to use the layout and coordinate information in the IMPACT ground truth XML files, and to obtain more detailed statistics on for instance frequent word errors, dictionary word hallucinations, dictionary coverage, …..

The evaluation data for each IMPACT language (Bulgarian, Czech, Dutch, English, French, German, Polish, Slovene, Spanish) consist of a random selection of about 200 pages from the “Evaluation” subset of the ground truth transcriptions.

5.1.1.1 Data used for the OCR evaluation

OCR evaluation is based on a dataset of approximately 200 pages for each language, selected randomly from the “Evaluation” subset of the demonstrator datasets. The appendix (section 11) lists the data used for each language.

5.1.1.2 Evaluation tool

We developed a special evaluation tool to collect information about in-dictionary errors (“false friends”, “dictionary hallucinations”), lexicon coverage, frequent errors, and to enable region-based alignment of OCR and ground truth. The version of the tool that was used in this report gives precision and recall for case-insensitive word accuracy, not counting punctuation, as the main evaluation metrics.

Most existing evaluation tools are plain text based. An additional benefit of the XML-based approach is that evaluation of specific parts of a document (headings, footnotes and, as we shall see in 5.1.1.3.4.1, named entities) is possible.

5.1.1.3 Results

The first results compare FineReader in its optimal internal dictionary and language setting (which is Old<SomeLanguage> whenever available, i.e. for English, French, German, Spanish) to FineReader with the same internal dictionary combined with an external dictionary - which in most cases seems to give the best results.

The following table summarizes the most important findings.

Language	*Default FineReader dictionary used*	*FineReader SDK version used*	*Recall with default dictionary*	*Precision with default dictionary*	*Recall with added historical lexicon*	*Precision with added historical lexicon*	*Absolute recall improvement*	*Relative recall improvement*[41]
Bulgarian	Bulgarian	9[42]	88,3	88,2	90,1	89,9	1,8	15,1
Czech	Czech	10	90,8	90,2	93,2	92,3	2,4	26,2
Dutch	Dutch	10	82,1	82,5	85,4	85,4	3,3	18,3
French	OldFrench	10	85,8	86	90	89,4	4,2	29,8
English	OldEnglish	10	83,2	83,2	82,6	82	-0,6	-3,6
English17	OldEnglish	10	84,2	85,5	87,6	88,2	3,4	21,4
German	OldGerman	10	73,5	76,9	77,5	80,4	4	15,1
Polish	Polish	10	30,3	32	37,3	37,7	7,1	10,1
Slovene	Slovene	10	74,1	73,6	82,1	81.0	8.0	30,9
Spanish	OldSpanish	10	71,7	76	76,1	79,5	4,4	15,6

5.1.1.3.1 Significance of results.

The table below gives the size of the OCR evaluation corpus in words, for each language.

Given the fact that the minimum corpus size is above 45000 words for all complete OCR evaluation sets, the amount of tokens is a sufficient sample for statistical significance in the sense that all Wilson confidence intervals for recall at alpha=0.05 are within a 1% range. Of course, this is not a realistic way to assess the predictive value of the measurements. It is obvious that our samples are extremely small when one considers the number of titles in the sample.

Language	*Corpus size*
Bulgarian	143018
Czech	61008

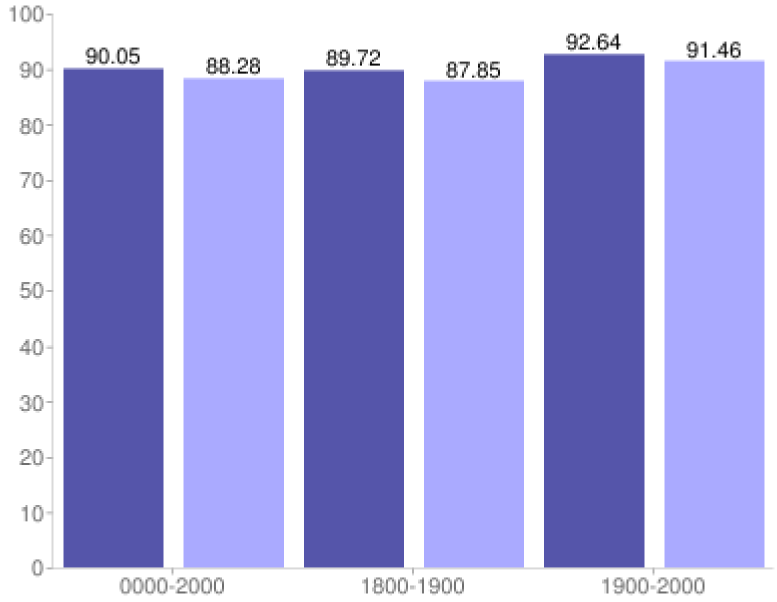
[41] (absolute recall improvement) / (100 – baseline recall)

[42] Bulgarian has been evaluated with FineReader 9, because we had technical problems combining external dictionaries with trained patterns in FR 10. Both the runs with and without external dictionary were executed with special trained user patterns for the obsolete characters, trained on a per-title basis.

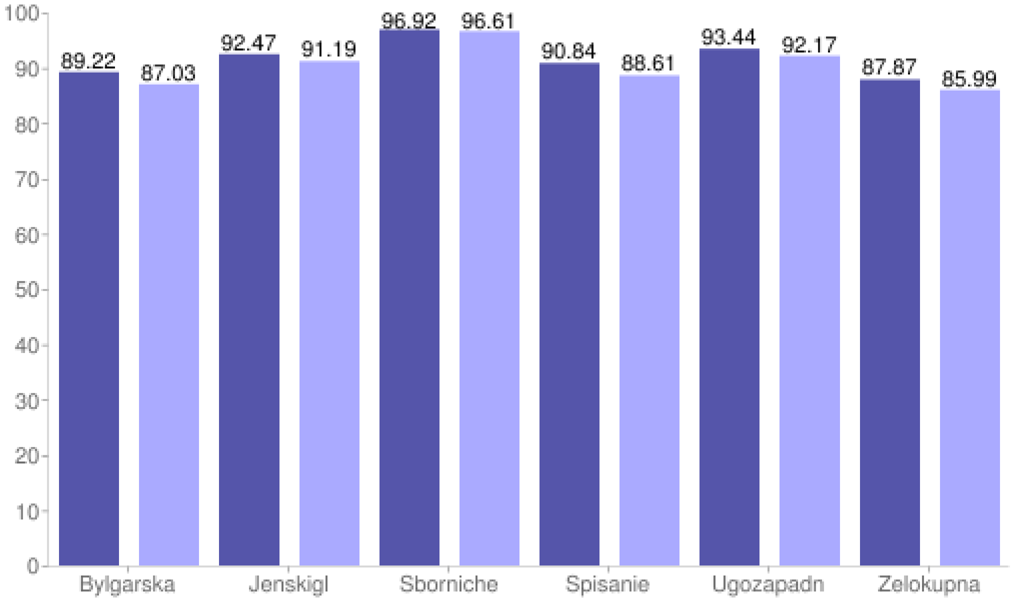
Dutch	343284
French	46281
English	92733
English17	49044
German	60372
Polish	67830
Slovene	45840
Spanish	45601

In the remaining part of this section, we shall briefly review results per language. In the charts, dark blue bars correspond to the word recall obtained with integration of historical lexica, whereas light blue bars give the performance of FineReader with its built-in dictionary. We will give the results per century and also per source. A detailed table with mean and standard deviation for each title is included to give an idea of the amount of variance.

5.1.1.3.2 Bulgarian



Word recal, per period



Word recall per title[43]

Performance is still slightly disappointing here, given the fact that the OCR distortion channel appears to be quite "colored" with a limited set of confusions causing a large proportion of the errors in the output of the FineReader engine with default dictionary:

GT→OCR	*frequency*	с→е	441
и→п	968	н→п	378
ѫ→ж	825	н→и	356
и→н	732	ъ→ь	270
п→н	579	ь→ъ	256

Of these, we find that some confusions are reduced considerably (с→е goes down to 200, ѫ→ж to 283), whereas others remain problematic (п→н still occurs 463 times, н→п 354 times). Dictionary coverage is 84,8 %, and "Dictionary word recall" (recall within the set of words in the lexicon) is 94,4%. When compared with other languages, the number of "false friends" (6,344 out of 14,352 recognition) is relatively low. The result may be influenced by the fact that we used the type-frequency list from the development ground truth as an OCR lexicon, and this still included some errors. (едпнъ: figures 22 times as a false friend; едннъ: 21 times).

Bylgarska iliustracia	28 pages	μ=0.886	σ=0.048
Jenski glas	29 pages	μ=0.925	σ=0.025
Sborniche za spomen na 25-godishninata ot smyrtta na Levski	6 pages	μ=0.966	σ=0.020
Spisanie Dennica	29 pages	μ=0.908	σ=0.049
Ugozapadna Bulgaria	10 pages	μ=0.935	σ=0.026
Zelokupna Bulgaria	23 pages	μ=0.886	σ=0.113
Distribution over all pages	125 pages	μ=0.908	σ=0.064
Distribution over titles:	6 titles	μ=0.918	σ=0.028

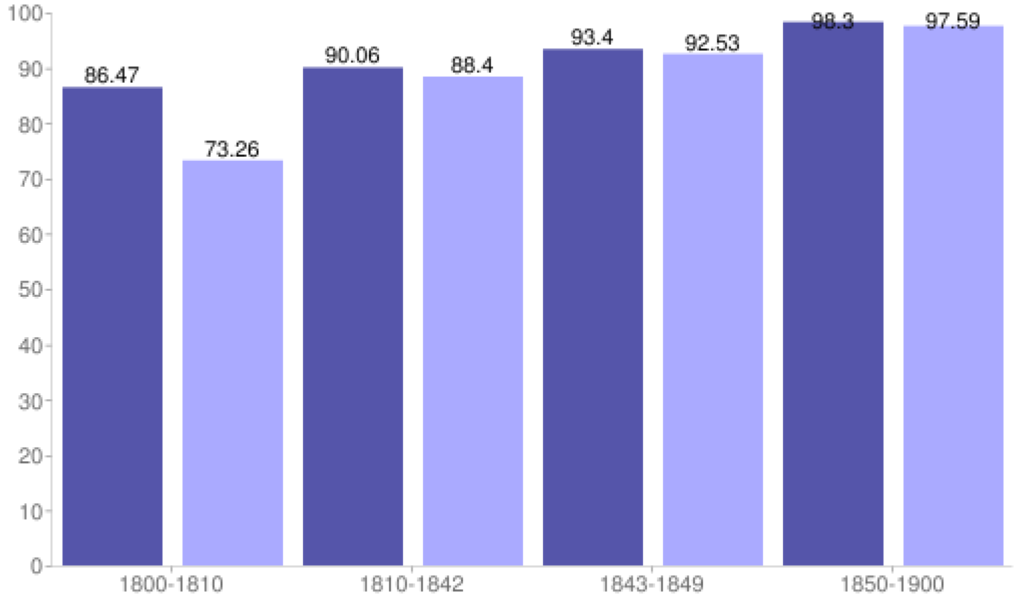
[43] For all languages, the list of titles can be found in the appendix

Mean and variance per title

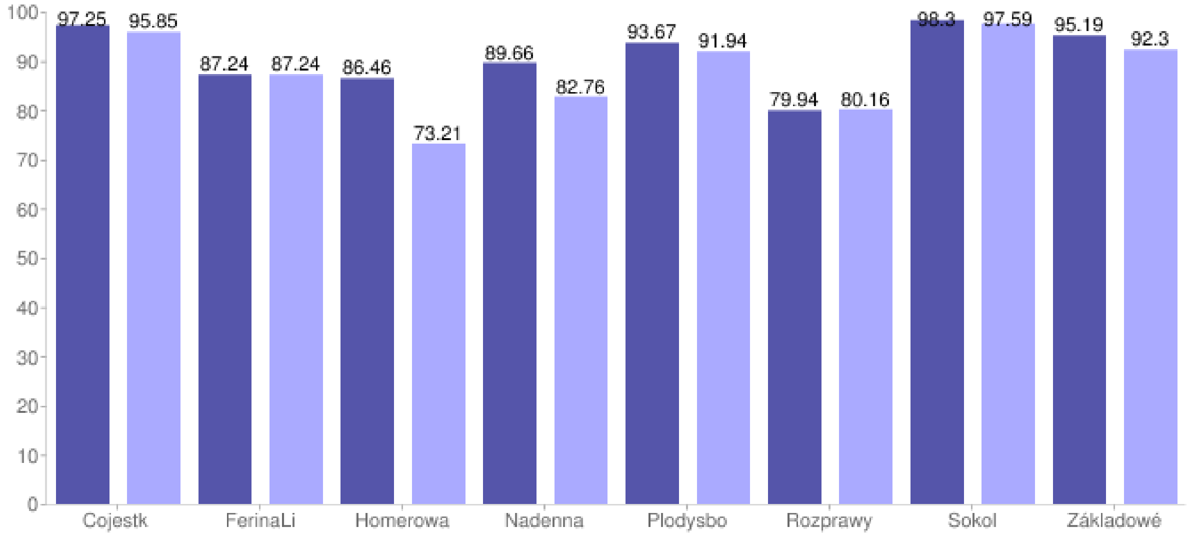
5.1.1.3.3 Czech

Czech obtains good results for all periods. This shows that a linguistically motivated list obtained by expansion of lemmata can sometimes do as well as a corpus-based list. Some remarks:

1. In this case, it pays off to split the dictionary into relatively small periods. We used separate external dictionaries for the four periods, as indicated by CUP. For German, split dictionaries also performed better than combined ones.
2. Czech is the only language in which already very good FineReader results are actually improved by the external dictionary. (Cf. the results for the *Sokol* journal)
3. For Czech, as for French and Spanish, many errors are related to the recognition of diacritics. These confusions are counted as "full errors".



Word recall per period



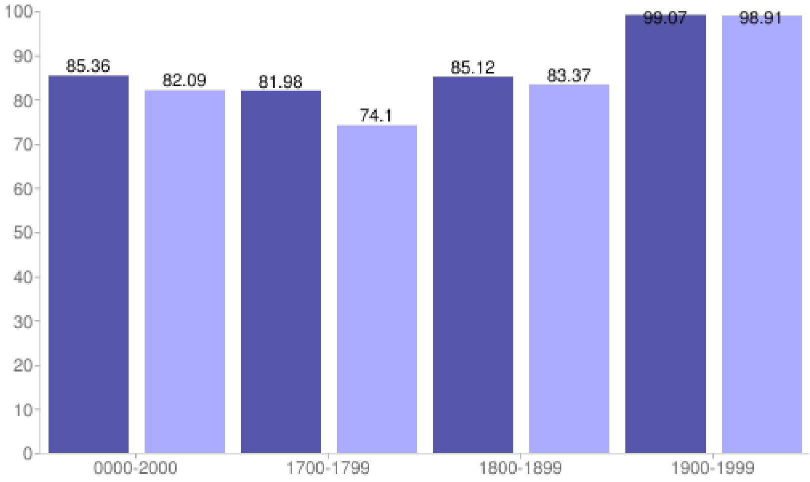
Word recall per title

Co jest konstituce?, čili, Krátký, prostonárodní wýklad hlawnějších zásad konstitucí ewropejských	28 pages	μ=0.973	σ=0.018
Ferina Lišák z Kuliferdy a na Klukově, čili, Kratičká historye zlopověstných kousků starého Reinecke	26 pages	μ=0.872	σ=0.084
Homerowa Iliada	23 pages	μ=0.859	σ=0.057
Na den narození neimocněišího, a neijasněišího cysare rímského, téz dědičného rakauského a krále ceského, Frantiska II., w Praze 12. den mesyce Unora, léta 1805	1 pages	μ=0.897	σ=0.000
Plody sborů učenců řeči českoslowanské prešporského	27 pages	μ=0.933	σ=0.048
Rozprawy o gmenách, počátkách i starožitnostech národu Slawského a geho kmeni /	27 pages	μ=0.787	σ=0.097
Sokol	28 pages	μ=0.878	σ=0.304
Základowé pitwy (Anatomie), čili, Soustawnj rozbor a popis těla lidského a gednotliwých geho částek	28 pages	μ=0.951	σ=0.022
Distribution over all pages	188 pages	μ=0.895	σ=0.143
Distribution over titles:	8 titles	μ=0.894	σ=0.055

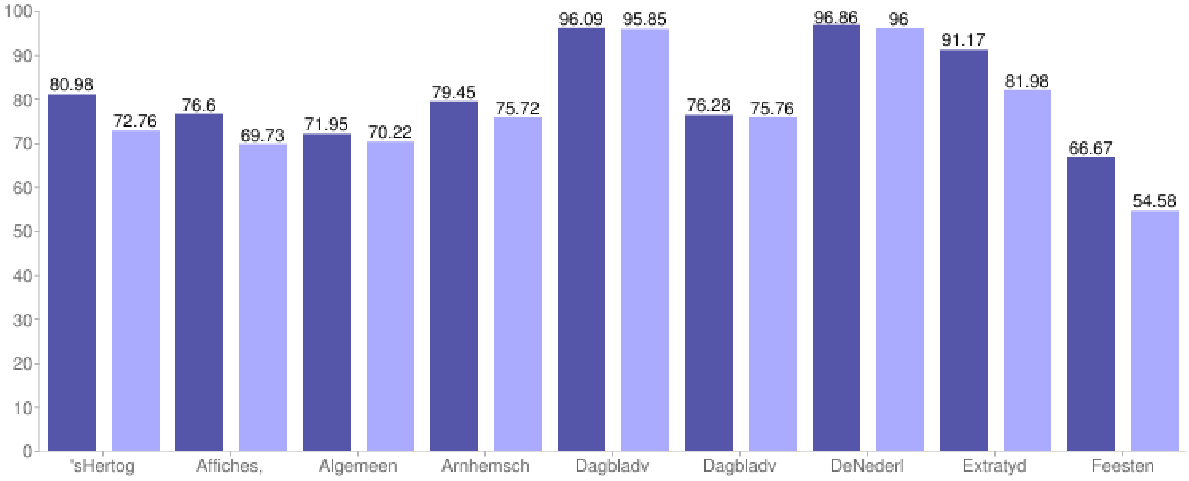
Mean and variance per title

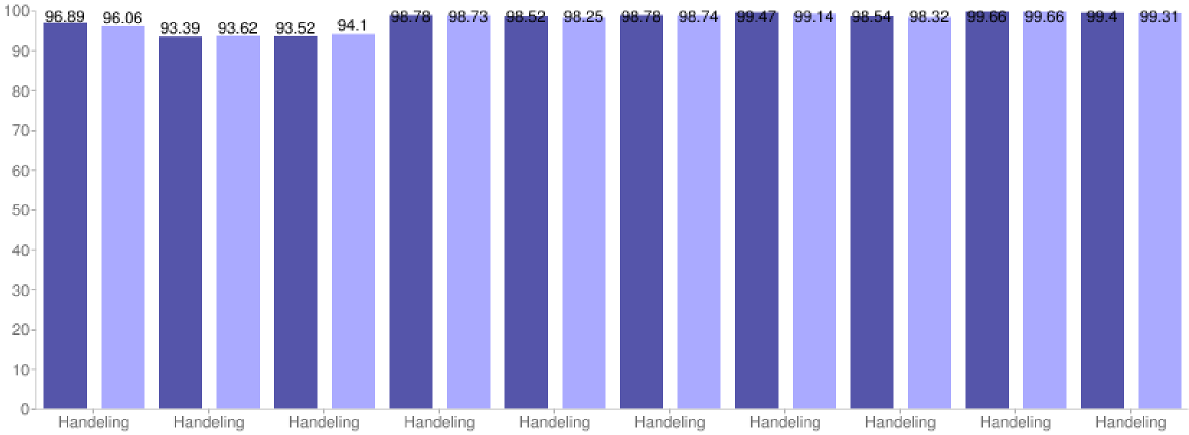
5.1.1.3.4 Dutch

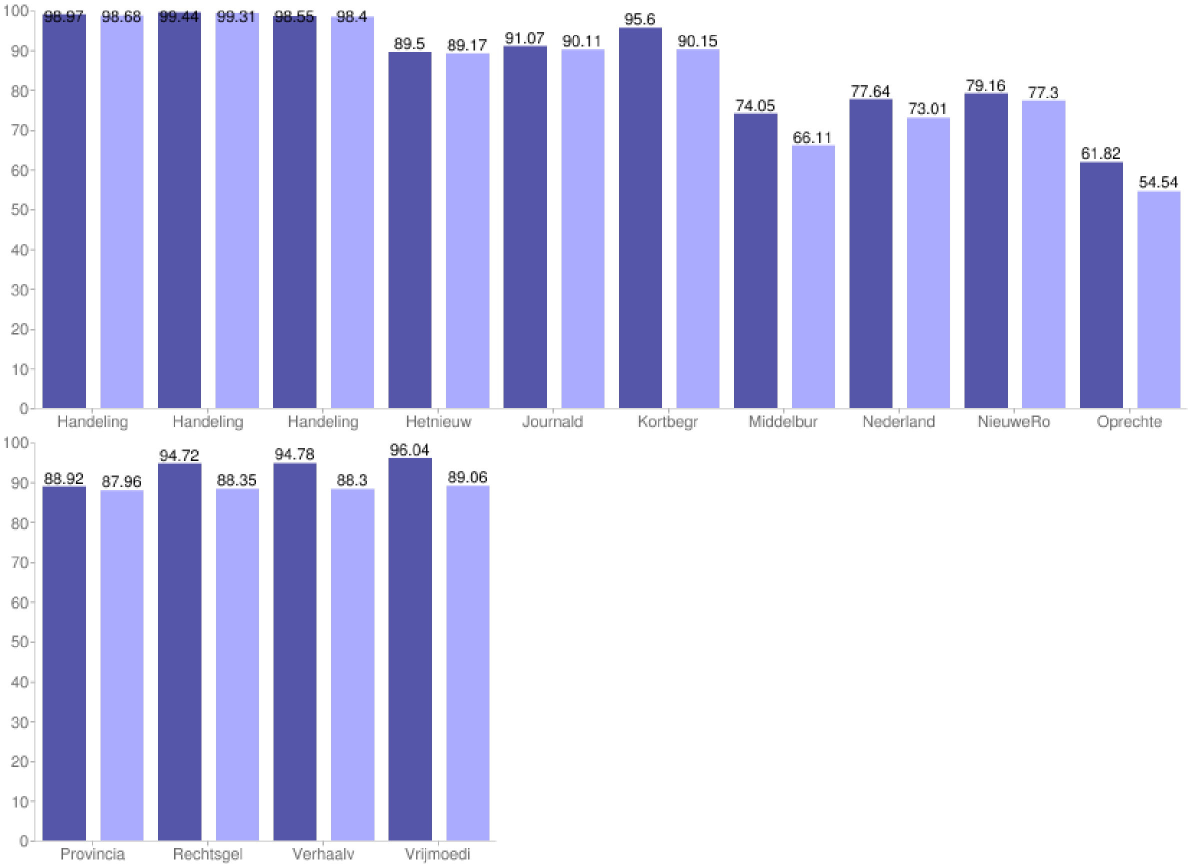
For Dutch, we use an automatically and cleaned version of the purely corpus-based word list because of its good coverage. Merging this with the cleaned lists we have been working on should improve performance, especially on the nineteenth and twentieth-century material. A significant part of the progress on eighteenth-century material is due to the “long s fix”, cf. 2.3.2.1).



Word recall per period







Word recall per title

's Hertogenbossche courant	24 pages	μ=0.815	σ=0.068
Affiches, annonces et avis divers d'Amsterdam = Advertentiën, aankondigingen en verschillende berigten van Amsterdam	2 pages	μ=0.752	σ=0.054
Algemeen Handelsblad	2 pages	μ=0.723	σ=0.058
Arnhemsche courant	4 pages	μ=0.763	σ=0.194
Dagblad van 's Gravenhage	1 pages	μ=0.961	σ=0.000
Dagblad van Zuidholland en 's Gravenhage	4 pages	μ=0.798	σ=0.117
De Nederlander : nieuwe Utrechtsche courant : (staatkundig- nieuws-, handels- en advertentie-blad) / onder red. van J. van Hall	2 pages	μ=0.969	σ=0.013
Extra tyding. Extract uit de resolutien der heeren Staaten van Holland [...] genomen op [...] 4. september 1786	5 pages	μ=0.730	σ=0.365
Feest en lydens stoffen voor de hervormde gemeente te Alkmaar	20 pages	μ=0.652	σ=0.228
Handelingen der Staten-Generaal 1826-1827	20 pages	μ=0.957	σ=0.027
Handelingen der Staten-Generaal 1857-1858	20 pages	μ=0.932	σ=0.022
Handelingen der Staten-Generaal 1868-1869	20 pages	μ=0.935	σ=0.019
Handelingen der Staten-Generaal 1891-1892	20 pages	μ=0.988	σ=0.003
Handelingen der Staten-Generaal 1929-1930	3 pages	μ=0.986	σ=0.005
Handelingen der Staten-Generaal 1931-1932	2 pages	μ=0.988	σ=0.006
Handelingen der Staten-Generaal 1932-1933	1 pages	μ=0.995	σ=0.000
Handelingen der Staten-Generaal 1933	1 pages	μ=0.985	σ=0.000
Handelingen der Staten-Generaal 1934-1935	1 pages	μ=0.997	σ=0.000
Handelingen der Staten-Generaal 1935-1936	4 pages	μ=0.994	σ=0.005
Handelingen der Staten-Generaal 1936-1937	2 pages	μ=0.990	σ=0.007
Handelingen der Staten-Generaal 1937-1938	4 pages	μ=0.994	σ=0.003
Handelingen der Staten-Generaal 1938-1939	2 pages	μ=0.985	σ=0.003
Het nieuws van den dag : kleine courant	9 pages	μ=0.887	σ=0.035
Journal de la province de Limbourg	5 pages	μ=0.913	σ=0.037

Kort begrip der waereld-historie voor de jeugd. / By J.F. Martinet	30 pages	μ=0.941	σ=0.083
Middelburgsche courant	18 pages	μ=0.739	σ=0.074
Nederlandsche staatscourant	6 pages	μ=0.783	σ=0.131
Nieuwe Rotterdamsche courant : staats-, handels-, nieuws- en advertentieblad	4 pages	μ=0.768	σ=0.114
Oprechte Haarlemse courant	14 pages	μ=0.622	σ=0.114
Provinciale Overijsselsche en Zwolsche courant : staats-, handels-, nieuws- en advertentieblad	4 pages	μ=0.886	σ=0.039
Rechtsgeleerd advis in de zaak van den gewezen stadhouder, en over deszelfs schryven aan de gouverneurs van de Oost- en West-Indische bezittingen van den staat [...]. Ingelevert [...] op den 7 january 1796. / By B. Voorda et al	20 pages	μ=0.948	σ=0.022
Verhaal van het levensgevaar, waar in zig drie Rotterdamsche burgers [...] bevonden hebben, te Utrecht	11 pages	μ=0.939	σ=0.045
Vrijmoedige aanmerkingen, over de uitsluiting van allen die door publieke armkassen bedeeld worden, als stemgerechtigden [...] bij eene oproeping van het Nederlandsche volk tot eene Nationaale Conventie	13 pages	μ=0.955	σ=0.024
Distribution over all pages	298 pages	μ=0.873	σ=0.149
Distribution over titles:	33 titles	μ=0.887	σ=0.112

Mean and variance per title

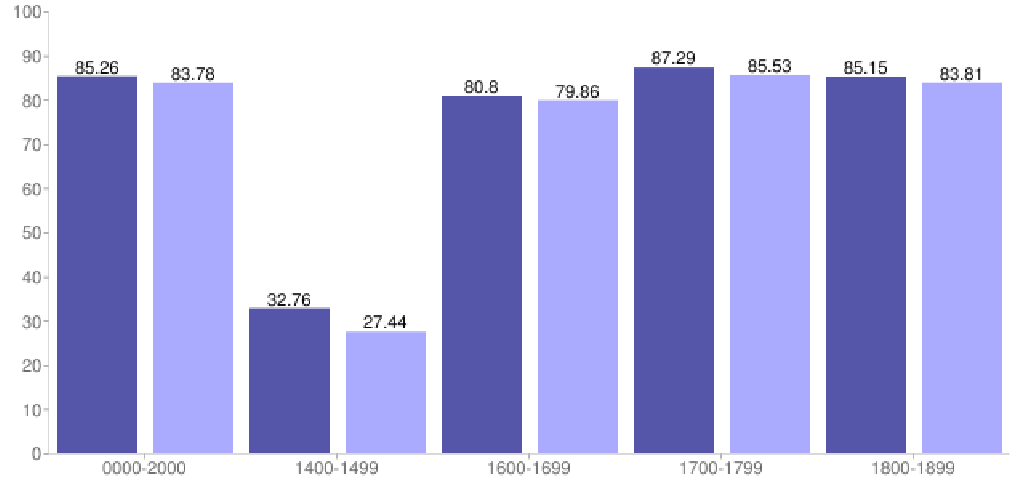
5.1.1.3.4.1 Evaluation of accuracy within Named Entities

For Dutch, we made a first start at evaluating the OCR of named entities by measuring the proportion of NE's which were recognized correctly. Measuring performance on NE's is possible because they are tagged with special tags in the INL dialect of Page XML, and the OCR evaluation tool can measure performance on specific parts of documents.

In this very limited evaluation, we reached a word recall of 90.1% for words tagged as part of named entities with the standard FineReader Dutch dictionary, and a word recall of 94.3% with the corpus-based external dictionary. We did not yet include the Named Entity lexicon developed for Dutch in this evaluation. A careful evaluation of the best way to combine the ‘general lexicon’ of appellatives with the NE lexica will be performed in the second Impact Extension. Evaluation of OCR performance of Named Entities is also foreseen for German.

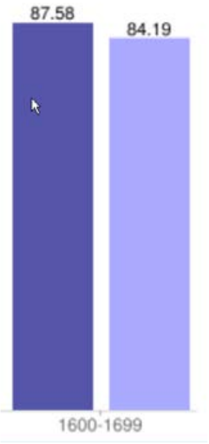
5.1.1.3.5 English

In assessing the (lack of) improvement achieved for English over the complete period covered, it should be born in mind that eighteenth-century English is already very close to modern, and that FineReader already provides dictionary support for older English. General tests were run using the complete lexicon (covering the whole IMPACT period). Results with a more fine-tuned lexicon will be presented in the next section.



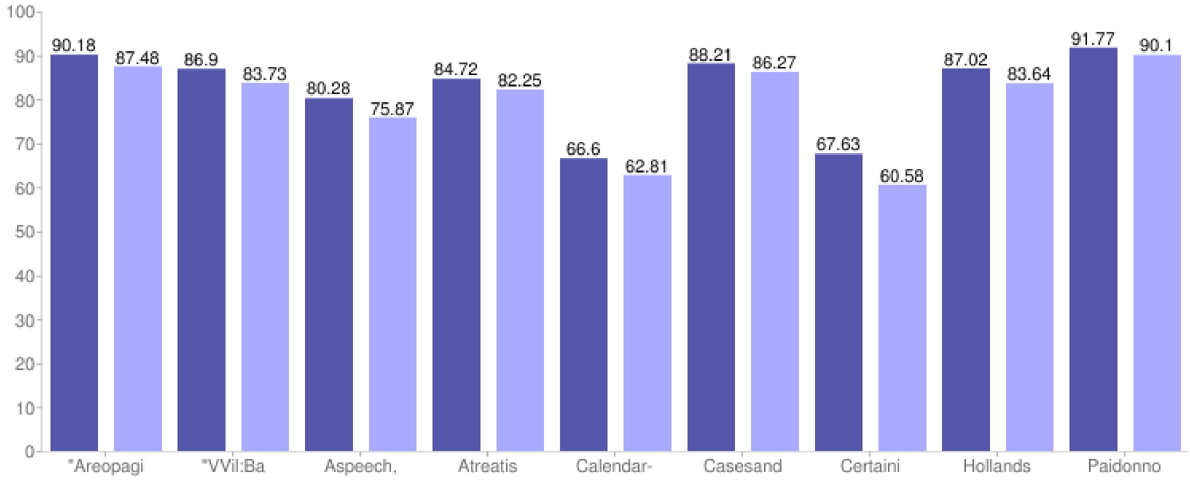
Word recall per period

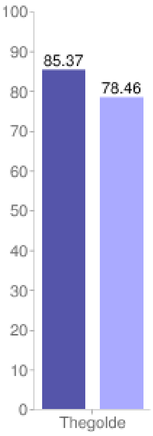
5.1.1.3.5.1 Focusing on seventeenth-century material



Various choices of lexicon (including a corpus-based one, using the ECCO corpus) did not bring about any significant improvement for eighteenth and nineteenth-century English. A more significant improvement is possible for seventeenth-century material, with a lexicon taken from OED quotations dating from 1580-1720.

Word recall per period





Word recall per title

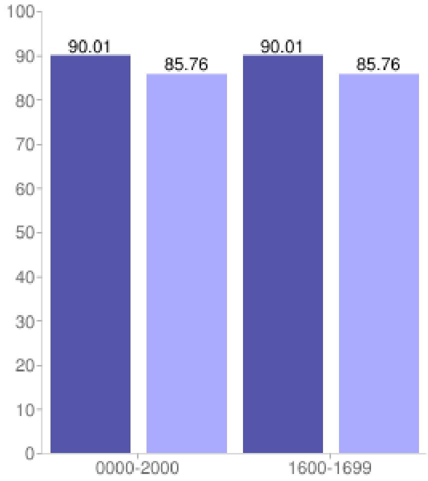
"Areopagitica; a speech of Mr. John Milton for the liberty of vnlicens'd printing, to the Parlament of England"	9 pages	μ=0.902	σ=0.026
"VVil: Bagnal's ghost. Or the merry devill of Gadmunton. In his perambulation of the prisons of London. / By E. Gayton, Esq;."	12 pages	μ=0.854	σ=0.045
A speech, or complaint, lately made by the Spanish embassadour to his Majestie at Oxford, upon	2	μ=0.806	σ=0.008

occasion of the taking of a ship called Sancta Clara in the port of Sancto Domingo, richly laden with plate, cocheneal and other commodities of great value, by	pages		
A treatise touching falling from grace. Or Thirteen arguments tending to prove that believers cannot fall from grace, as they were laid down at a conference at Yalding in Kent, examined and answered, with many absurdities of that doctrine shewed. Whereunt	10 pages	µ=0.841	σ=0.040
Calendar-reformation. Or, An humble addresse to the Right Honorable the Lords and Commons assembled in Parliament, touching the dayes and moneths, that they may be taught to speak such a language as may become the mouth of a Christian. / By I.B.	2 pages	µ=0.660	σ=0.009
Cases and questions resolved in the civil-lavv. Collected by R. Zouch professor of the civil-law in Oxford.	62 pages	µ=0.881	σ=0.053
Certain information from Devon and Dorset: concerning the Commission of Array.	1 pages	µ=0.676	σ=0.000
Hollands ingratitude, or, A serious expostulation with the Dutch shewing their ingratitude to this nation, and their inevitable ruine, without a speedy compliance and submission to His Sacred Majesty of Britain / by Charles Molloy of Lincolns-Inn, Gent.	9 pages	µ=0.872	σ=0.039
Paidon nosemata· = or Childrens diseases both outward and inward. From the time of their birth to fourteen years of age. With their natures, causes, signs, presages and cures. In three books: 1. Of external 2. Universal 3. Inward diseases. Also, the resol	47 pages	µ=0.915	σ=0.046
The golden trade: or, A discouery of the riuer Gambra, and the golden trade of the Aethiopians Also, the commerce with a great blacke merchant, called Buckor Sano, and his report of the houses couered with gold, and other strange obseruations for the good	37 pages	µ=0.851	σ=0.053
Distribution over all pages	191 pages	µ=0.876	σ=0.061
Distribution over titles:	10 titles	µ=0.826	σ=0.084

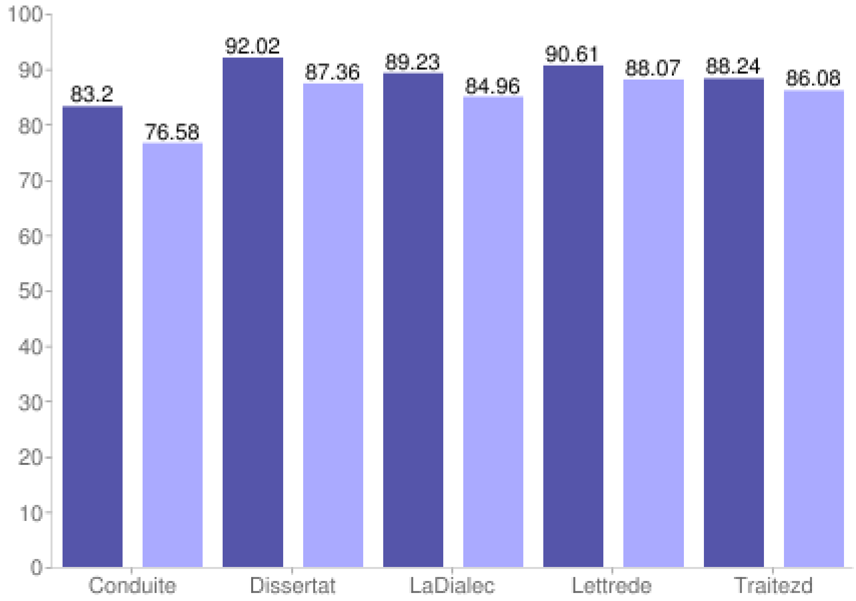
Mean and variance per title

5.1.1.3.6 French

The French OCR lexicon performs well in terms of error reduction. This may be partly due to the fact that all texts chosen for IMPACT are of the same type and belong to the philosophical subject domain. The simple explanation that this might be due to a good coverage is not entirely straightforward as we do not know the content of the internal lexicon.



Word recall per period



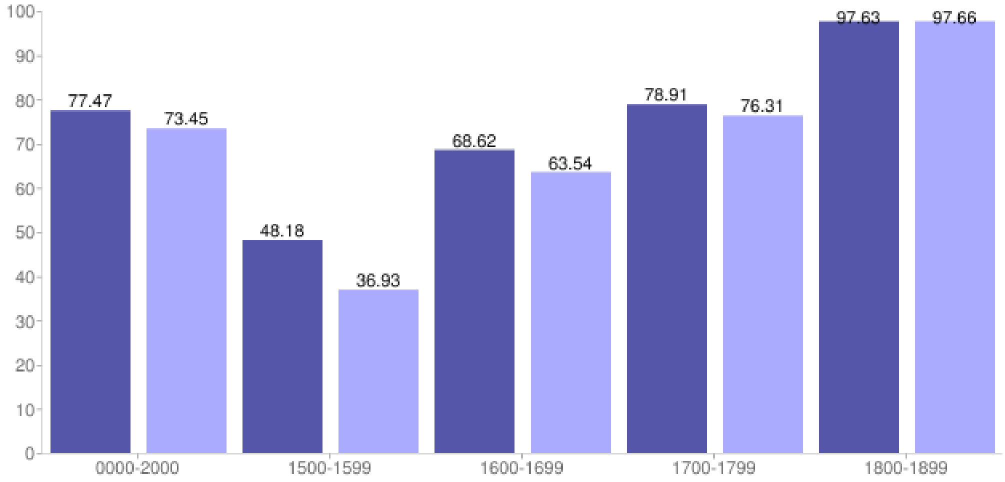
Word recall per title

Conduite du jugement naturel où tous les bons esprits de l'un et l'autre sexe pourront facilement puiser la pureté de la science, par M. Jacques Forton, sieur de S. Ange,...	38 pages	μ=0.813	σ=0.158
Dissertation de la philosophie en général	40 pages	μ=0.920	σ=0.020
La Dialectique du sieur de Launay, contenant l'art de raisonner juste sur toute sorte de matières...	39 pages	μ=0.890	σ=0.046
Lettre de M. Gadroys à M. de La Grange Trianon,... pour servir de réponse à celle que M. de Castelet a écrite contre les raisons de M. Descartes touchant le flux et le reflux de la mer. - Seconde lettre de M. Gadroys... [au même, sur le même sujet.]	21 pages	μ=0.909	σ=0.043
Traitez de métaphysique démontrée selon la méthode des géomètres. [Par le sieur de La Coudraye.]	39 pages	μ=0.881	σ=0.053
Distribution over all pages	177 pages	μ=0.881	σ=0.090
Distribution over titles:	5 titles	μ=0.883	σ=0.037

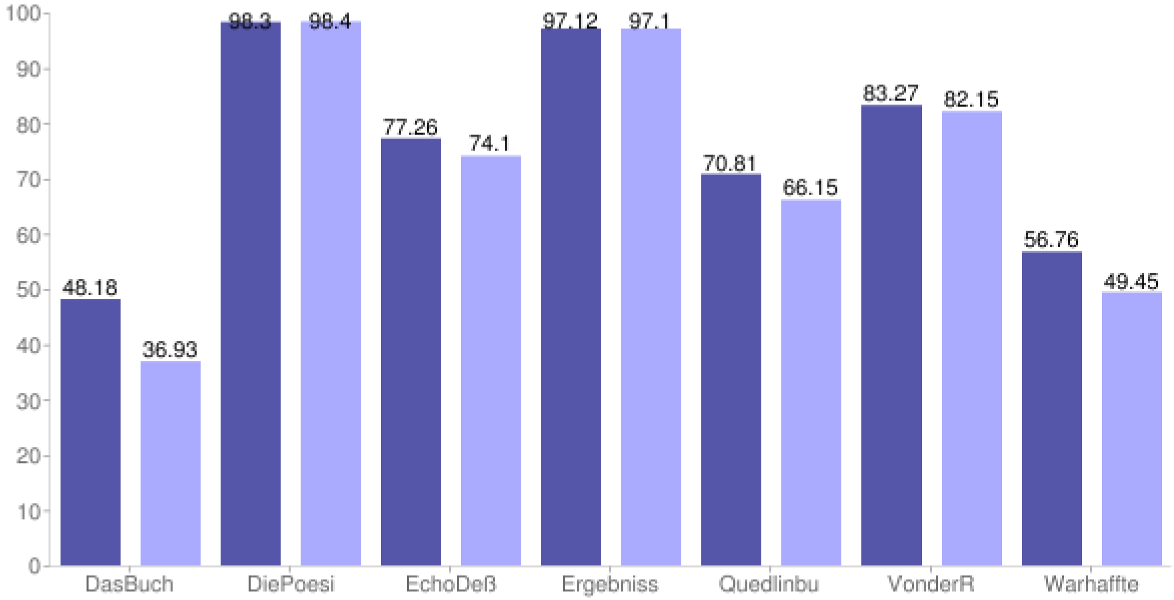
Mean and variance per title

5.1.1.3.7 German

For German, as for Dutch, we see significant improvement mainly for the older texts. Another observation is that special cleaning of the OCR lexica and splitting them on a per-century basis improved results. The results for the sixteenth and seventeenth century will be re-evaluated with the most recent FineReader version, which incorporates better recognition for Gothic fonts.



Word recall per period



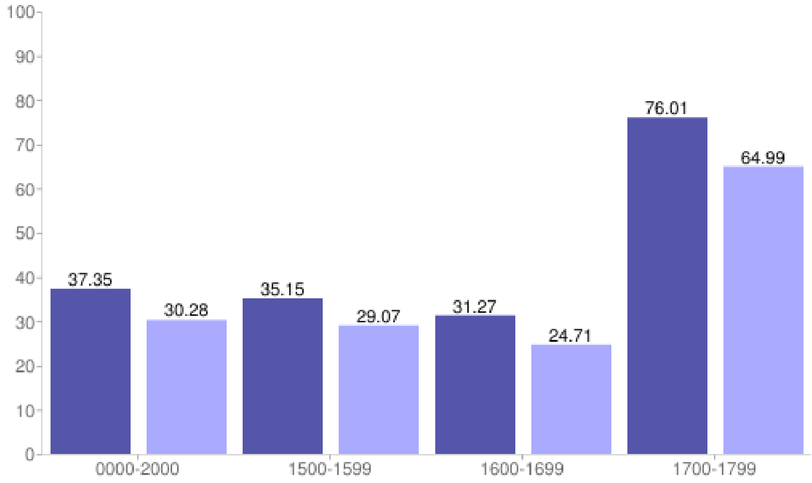
Word recall per title

Das Buch des heyligen Römischen Reichs unnderhalltunge	41 pages	μ=0.457	σ=0.113
Die Poesie ihr Wesen und ihre Formen mit Grundzügen der vergleichenden Literaturgeschichte	25 pages	μ=0.983	σ=0.010
Echo Deß Hochzeitlichen Te Deum Laudamus	17 pages	μ=0.757	σ=0.079
Ergebnisse der Erhebungen über die Beschäftigung gewerblicher Arbeiter an Sonn- und Festtagen, Bd.:1, Gruppe I bis VII der Gewerbestatistik, Berlin, 1887	25 pages	μ=0.969	σ=0.020
Quedlinburgisches Kreis-Tags-Memorial	37 pages	μ=0.700	σ=0.081
Von der Regierung der Kirche und den unterschiedlichen Würden der Geistlichkeit *(full title in comments)	15 pages	μ=0.833	σ=0.041
Warhaffter und grundlicher Bericht uß was Ursachen Martinus du Voysin (zu Basel verburgerter Krämer) inn der Statt Surseew im Aargöw, ..., den 13. Tag Octobris deß 1608. Jars erstlich enthauptet, und volgends verbrennt worden	9 pages	μ=0.569	σ=0.060
Distribution over all pages	169 pages	μ=0.733	σ=0.210
Distribution over titles:	7 titles	μ=0.753	σ=0.181

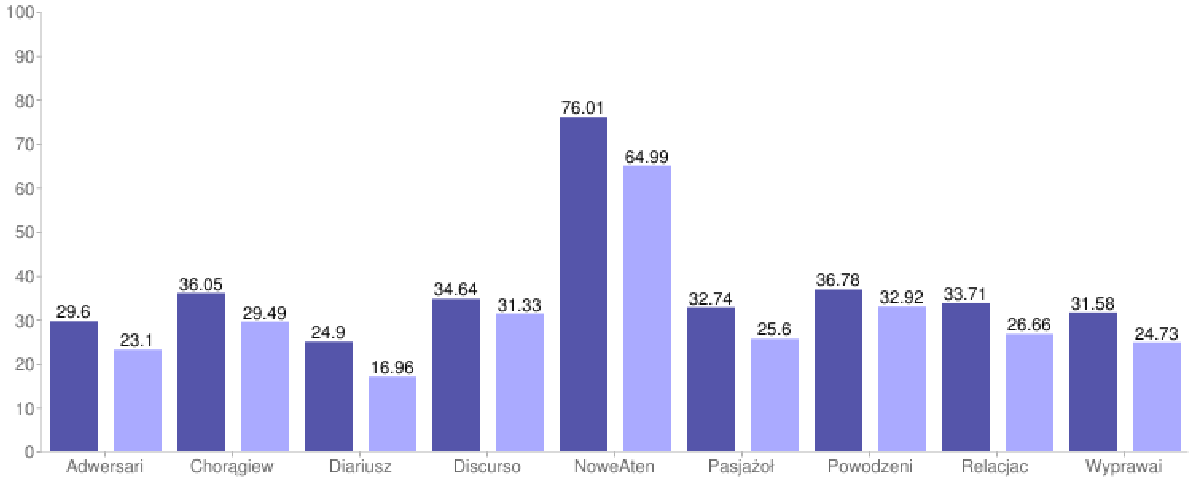
Mean and variance per title

5.1.1.3.8 Polish

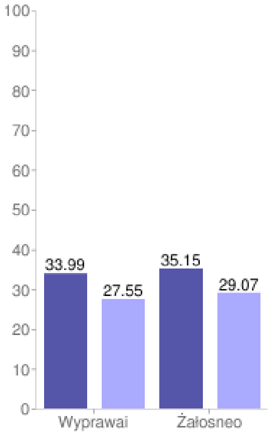
The Polish OCR lexicon leads to improvement of text recognition, but cannot compensate for the fact that FineReader is not trained for early black-letter Gothic print[44]. The same phenomenon wrecks the performance for early German, Dutch and English (the *Prologus*), even in cases where the print and image quality seem good. A somewhat more usable improvement is realized on eighteenth-century material, which mainly consists of the *Nowe Ateny* encyclopedia.



Word recall per period



[44] The recent improvements implemented in FineReader 10 for earlier Gothic fonts will be evaluated in the second IMPACT extension.



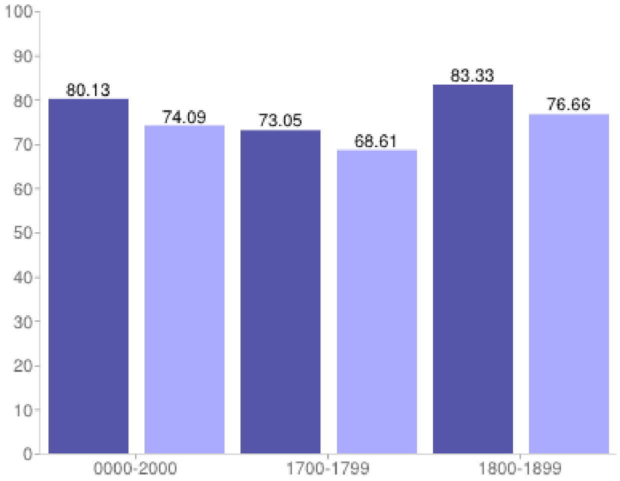
Word recall per title

Adwersaria, albo terminata sprawy wojennej, która się toczyła w wołoskiej ziemi z tureckim cesarzem	24 pages	μ=0.310	σ=0.073
Chorągiew Sarmacka w Wołoszech, to jest pospolite ruszenie i szczęśliwy powrót Polaków z Wołoch w roku 1621	11 pages	μ=0.356	σ=0.055
Diariusz wiadomości od wyjazdu króla z Wilna do Smoleńska	29 pages	μ=0.248	σ=0.066
Discurs o cenie pieniedzy teraznieyszey y o niektorych skutkach iey…	22 pages	μ=0.346	σ=0.046
Nowe Ateny, albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... . Część 3 albo Supplement.	29 pages	μ=0.761	σ=0.047
Pasja żołnierzy obojga narodów w stolicy moskiewskiej krótko opisana	16 pages	μ=0.322	σ=0.066
Powodzenia niebezpiecznego ale szczęśliwego wojska j. k. m. w Multanach opisanie	6 pages	μ=0.355	σ=0.036
Relacja chwalebnej ekspedycji Jana Kazimierza, króla polskiego i szwedzkiego	23 pages	μ=0.341	σ=0.034
Wyprawa i wyjazd sułtana Amurata, cesarza tureckiego, na wojnę do Korony Polskiej	29 pages	μ=0.321	σ=0.064
Wyprawa i wyjazd sułtana Amurata, cesarza tureckiego, na wojnę do Korony Polskiej_BW	29 pages	μ=0.344	σ=0.068
Żałosne opisanie upadku króla hiszpańskiego na morzu i na lądzie	29 pages	μ=0.350	σ=0.031
Distribution over all pages	247 pages	μ=0.376	σ=0.155
Distribution over titles:	11 titles	μ=0.369	σ=0.128

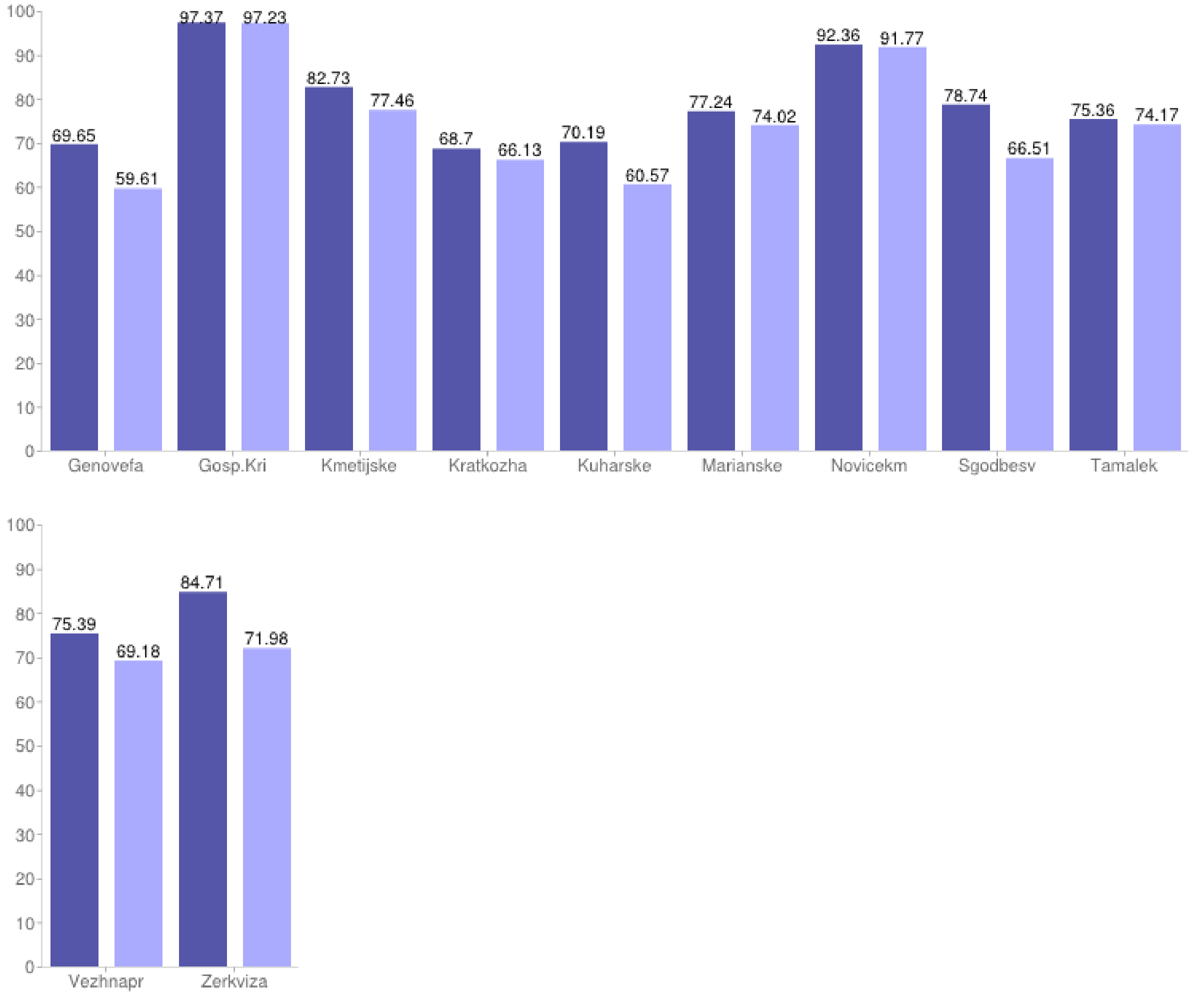
Mean and variance per title

5.1.1.3.9 Slovene

The best results as to relative word recall improvement have been obtained for Slovene.



Word recall per period



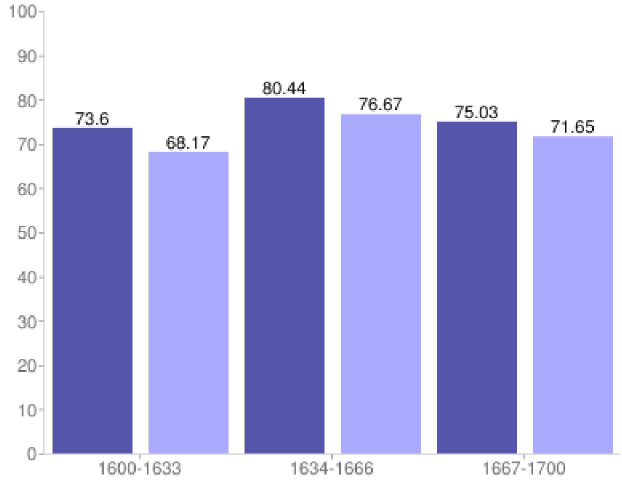
Word recall per title

Genovefa	20 pages	μ=0.71	σ=0.03
Gosp. Krištofa Šmida korarja avgustanskiga, zgodBe S. Pisma za mlade ljud...	20 pages	μ=0.97	σ=0.02
Kmetijske in rokodelske novice	11 pages	μ=0.82	σ=0.11
Kratkozhasne uganke	19 pages	μ=0.71	σ=0.12
Kuharske Bukve	19 pages	μ=0.71	σ=0.06
Marianske Kempensar, ali Dvoje bukuvze	20 pages	μ=0.81	σ=0.05
Novice kmetijskih, rokodelnih in narodskih reči	3 pages	μ=0.92	σ=0.03
Sgodbe svetiga pisma za mlade ljudi	20 pages	μ=0.82	σ=0.05
Ta male katechismus	19 pages	μ=0.82	σ=0.05
Vezhna pratika od gospodarstva	19 pages	μ=0.77	σ=0.05
Zerkviza na skali	14 pages	μ=0.85	σ=0.06
Distribution over all pages	184 pages	μ=0.80	σ=0.1
Distribution over titles:	11 titles	μ=0.81	σ=0.08

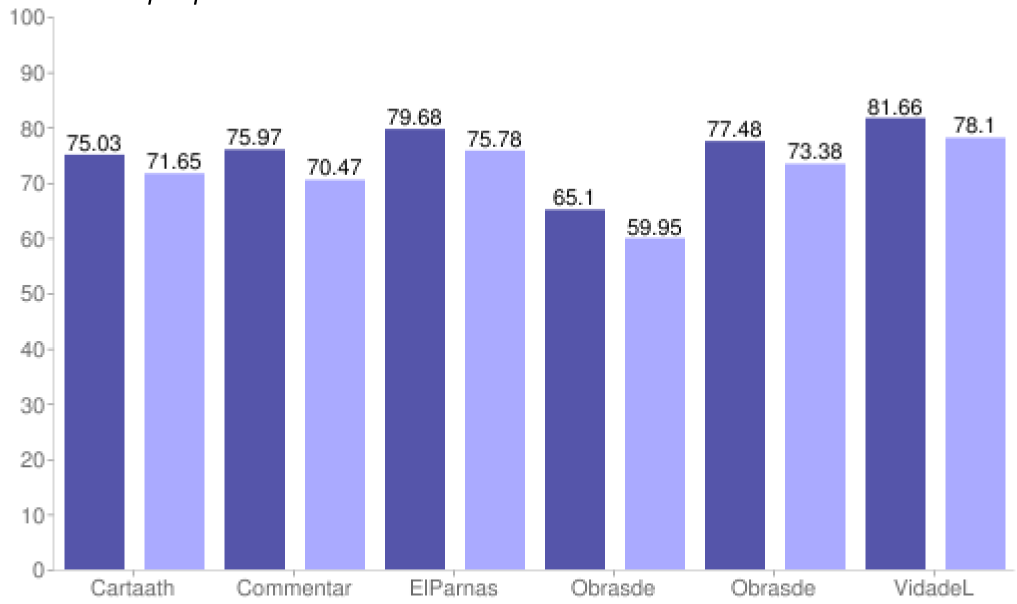
Mean and variance of per page scores per title

5.1.1.3.10 Spanish

For Spanish, we used the words from the *type_frequencies* table of the lexicon database submitted by the University of Alicante team. This means that the words from the background development corpus for the IR lexicon work is used as an OCR dictionary. As for French, some suspicious words seemed to hinder performance using the first pre-release of the Spanish lexicon. We manually removed a few frequent incorrect word forms, like "quc", which already improved performance by 1-2%. As for French, the method of removing infrequent short words[45] improved performance for Spanish by a few percent.



Word recall per period



Word recall per title

Carta athenagorica	33 pages	µ=0.751	σ=0.092
Commentarios reales	33 pages	µ=0.760	σ=0.055
El Parnasso español	33 pages	µ=0.789	σ=0.076
Obras de Garcilasso de la Vega con las anotaciones por el Mtro. Francisco Sánchez Brocense	33 pages	µ=0.650	σ=0.097
Obras de Lope de Vega	33 pages	µ=0.782	σ=0.107
Vida de Lazarillo de Tormes	33 pages	µ=0.810	σ=0.069
Distribution over all pages	198 pages	µ=0.757	σ=0.099
Distribution over titles:	6 titles	µ=0.757	σ=0.052

[45] Cf. section 2.3.3.1

Mean and variance per title

5.1.2 Conclusions and possibilities for further development

Results from the addition of special historical lexica show improvement ranging from about 10% to about 30% of word error reduction, relative to the total number of word errors. Contrary to what was expected, it is difficult to predict the added value of the external dictionary in terms of initial OCR accuracy or coverage of the historical lexicon. This is partly due to the fact that the content of the FineReader internal dictionaries is not known to us. This leads to unexpected results, like for instance when comparing French with Bulgarian. For both languages the coverage of the IMPACT lexicon is around 80% on the evaluation set, but for French the improvements are much more significant than for Bulgarian.

Another hypothesis that was not supported by the results was that extremely poor quality OCR would never show improvement using an added historical lexicon, although using an added historical lexicon by no means resolves all issues.

The assumption, finally, that adding a type frequency list without further cleaning would suffice as a good OCR lexicon also proved to be wrong. Results and initial experiments have indicated that there is room for improvement as to the content of an OCR lexicon.

There are several possibilities to enhance the quality of the OCR. Below, we will describe what strategies we will focus on in the second IMPACT extension. Results thereof will be integrated in a new version of this paper.

5.1.2.1 Further refinement of recipes for good OCR lexicon content

Finding an optimal lexicon for a given dataset turns out to be no trivial task, as discussed in section 2.3.3. Further development of good heuristics for transforming a corpus-based type-frequency list into an OCR lexicon in terms of weights and which words to include, is a prerequisite for arriving at a true “CookBook” for OCR lexica. Some experiments are scheduled for 2012 in the second IMPACT extension.

5.1.2.2 Deployment of hypothetical lexica and morphological analysis

In order to improve recognition of out-of-vocabulary words, a module for the integration of two-level finite-state morphology with the help of the XSFT tools has been developed using the FineReader external dictionary interface. However, evaluation of this approach requires setting up a suitable test case and preparing the necessary data, which has not yet been done.

5.1.2.3 Testing with other OCR engines

In the second IMPACT extension, the trainable open-source OCR tesseract will be evaluated, notably/particularly in the context of Polish historical data. We hope this will also provide some insights into lexicon deployment for this engine.

5.2 Profiler experiments and evaluation for several languages

This section will be added by LMU in the course of the second IMPACT extension.

6 Deployment and evaluation of IR lexica

6.1 Deployment of IR lexica

As has been explained before, the main reason to develop lexica with linguistic information instead of just gathering word lists from corpora is that besides contributing to better OCR, such a lexicon can be used to improve IR.

As is the case for OCR, linguistic data can be deployed in many ways, with varying degrees of sophistication, to improve retrieval. Part-of-speech tagging can be used to improve precision; recall could be improved by applying approximate matching procedures. In this section, we shall try to assess the value of the lexica produced by IMPACT by means of a simple two-step matching procedure to relate the modern standard dictionary form of a word to its actual occurrences in historical text.

The procedure supposes the availability of the following resources: a historical lexicon which enables us to look up the dictionary headword ("(modern) lemma[46]") form for historical word forms, a modern lexicon which does the same for modern forms, and a set of *patterns* which link historical to modern spelling. So there are two distinct ways of arriving at the modern lemma form *chancellor* from the historical word form *chaunceleres.*

Either the word is listed in the lexicon, in which case we simply have:

1: Historical word form → (consultation of historical lexicon): Modern lemma
chaunceleres *chancellor*

As a fall-back (supposing the word is not in the historical lexicon), we can try to match the spelling of the unknown historical word to a known modern word:

2: Historical word form → Corresponding modern word form[47] → Modern lemma
chaunceleres - - (matching) - - chancellors - - (lemmatization) chancellor

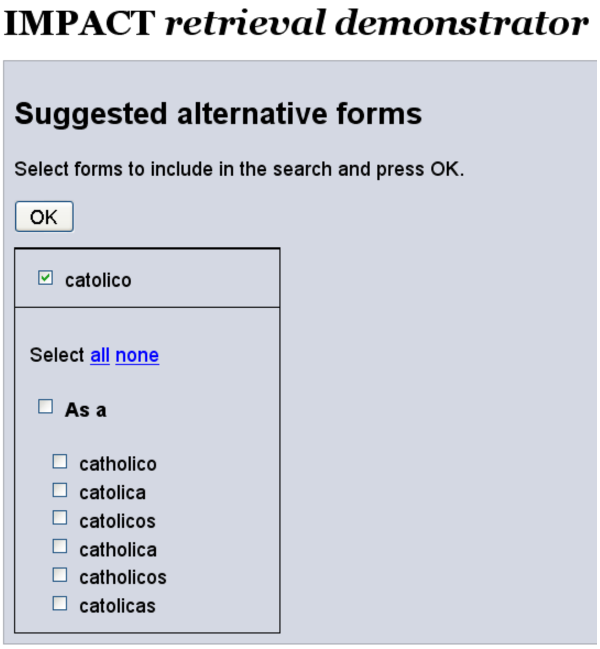
Because in IMPACT we deal with documents that are linguistically close to modern, it also makes sense to simply add the modern lexicon for consultation:

3: Historical word form = modern word form → Modern lemma

There are several ways in which this procedure can be at least partially integrated into real IR applications to enhance recall. The lexica (without the pattern matching) and/or the patterns can be used for query expansion. Alternatively, text could be (semi)-automatically pre-lemmatized before indexing, or the pattern-matching step could be reversed to match a modern word against the list of historical words found in the corpus.

[46] In IMPACT lexica, the lemma for each entry is always in modern spelling

[47] In IMPACT, we only apply the pattern matching between modern and historical word forms, but not between historical word forms found in the historical lexicon and historical word forms found in the corpus. This is based on the observation that we often already have a good full-form lexicon for the modern language, and high-coverage historical lexica are much harder to come by. The exception is English, where we do not use a modern lexicon at all and apply the pattern matching to the historical lexicon.



IMPACT *retrieval demonstrator*

Suggested alternative forms

Select forms to include in the search and press OK.

☑ catolico

Select all none

□ As a

□ catholico
□ catolica
□ catolicos
□ catholica
□ catholicos
□ catolicas

Figure: suggestions for historical variants of the query can be presented to the user before query expansion

6.2 Evaluation

6.2.1 Data used for IR evaluation

For each of the languages in IMPACT, the following resources are needed:

1. An IR lexicon for historical language
2. A modern lexicon
3. A set of spelling variation rules, linking historical to modernspelling
4. An evaluation set of a limited (~10.000) amount of tokens of running text, annotated with modern lemma or modern word form, or both.

We were lucky to have these resources available in time for all languages. The main exception is that we did not use a modern lexicon for English[48]. As to lemma vs. modern word form annotation, all evaluation sets have modern lemma, with the exception of the “modern” words in the German data, and Bulgarian, French, German and Slovene annotated modern word forms. Since 2-4 are IMPACT deliverable resources, described in this report and in the deliverable documentation, we shall briefly review the modern lexica used.

- For Bulgarian, we received a lexicon of 497,421 word forms from BAS, containing 46,170 distinct lemmata.
- For Czech, we used a modern-spelling lexicon from CUP, based on the same set of lemmata that was used for the historical lexica.
- For Dutch, we used a combination of JVKLeX, a relatively small internal modern lexicon produced at the INL, and the results of automatic reverse lemmatization of the set of WNT modern lemmata.
- For English, no modern lexicon was used.
- For French, we had Morphalou available.

[48] Since the Oxford English Dictionary is very extensive both as a historical and as a contemporary resource, including an external Modern English full-form lexicon would probably hardly improve the results.

- For German, we used the Morphy[49] German lexicon. In the experiments in Gotscharek et al., the more comprehensive CISLeX was used.
- For Polish, we used the *grammatical dictionary* data (cf. section 3.7.4).
- For Slovene, a lexicon based on the Slovene MULTEXT-East lexicon was used.
- For Spanish, the *Apertium* lexicon was used (cf. 3.9.4).

6.2.2 IR evaluation

A true IR evaluation[50] - in the sense of assessment the degree to which a user's information need is satisfied by a retrieval result - requires a manually annotated collection of relevance judgments, which is not feasible in the context of this study. Instead, in keeping with the general IMPACT objective, which is to do ground work for better IR rather than to try to produce an optimal search engine for historical text, we merely evaluate the results of applying the matching procedure described above as a first step towards better retrieval in historical documents.

The IR experiment is set up as follows:

- The user is conceived to be looking for tokens in the evaluation corpus, specifying a modern lemma form as a query (for instance *meet)*
- The system delivers a set of matching tokens according to the procedure sketched in 6.1.
- Correctness of the query results depends on the correctness of the matching procedure in each context. The *true* set of retrieval results is taken to be the set of tokens corresponding to the specified lemma. Note that this penalizes ambiguity within the lexicon: *meeting* in "*a fruitful meeting*" would be considered an incorrect match.

6.2.3 Corpus-linguistic evaluation

From the perspective of a corpus linguist, one could simply evaluate the degree to which we are successful in assigning the lemma form to tokens in historical text as a metric. As a way to estimate recall, this is indeed more or less satisfying. Bear in mind that the matching procedure assigns several lemma candidates to each corpus token, so simply *lemmatization accuracy = number of correctly lemmatized words / total number of words* is not an option.

Thus, as a first approximation,

$$recall = \frac{number\ of\ corpus\ tokens\ with\ at\ least\ one\ correct\ lemma\ candidate}{number\ of\ \ tokens\ in\ corpus}$$

$$precision = \frac{number\ of\ correct\ lemma\ assignments\ in\ the\ corpus}{total\ number\ of\ \ lemma\ assignments}$$

$$\text{F1} \quad =2 * \frac{precision*recall}{precision+recall}$$

This means that for a corpus of 10,000 tokens, in which the matching procedure returns on average two candidates per token (so 20,000 lemmatizations have been suggested by the matching procedure) and 8000 tokens have at least one correct lemma suggestion, recall is 80% (8000/10,000) and precision is 8000 / 20,000 = 40%, and the F1 score, combining precision and recall in one metric, equals 2 * 0.8*0.4 / 1.2 = 0.64/1.2 = 0.53.

We take a slightly different path compared with (Gotscharek et al. 2010), where the second step, lemmatization, which further increases the ambiguity already inherent in the matching step, is not taken into account. We need not discuss in detail whether it is more relevant to evaluate modern word form assignment or modern lemma assignment. It could be argued that evaluation on the word

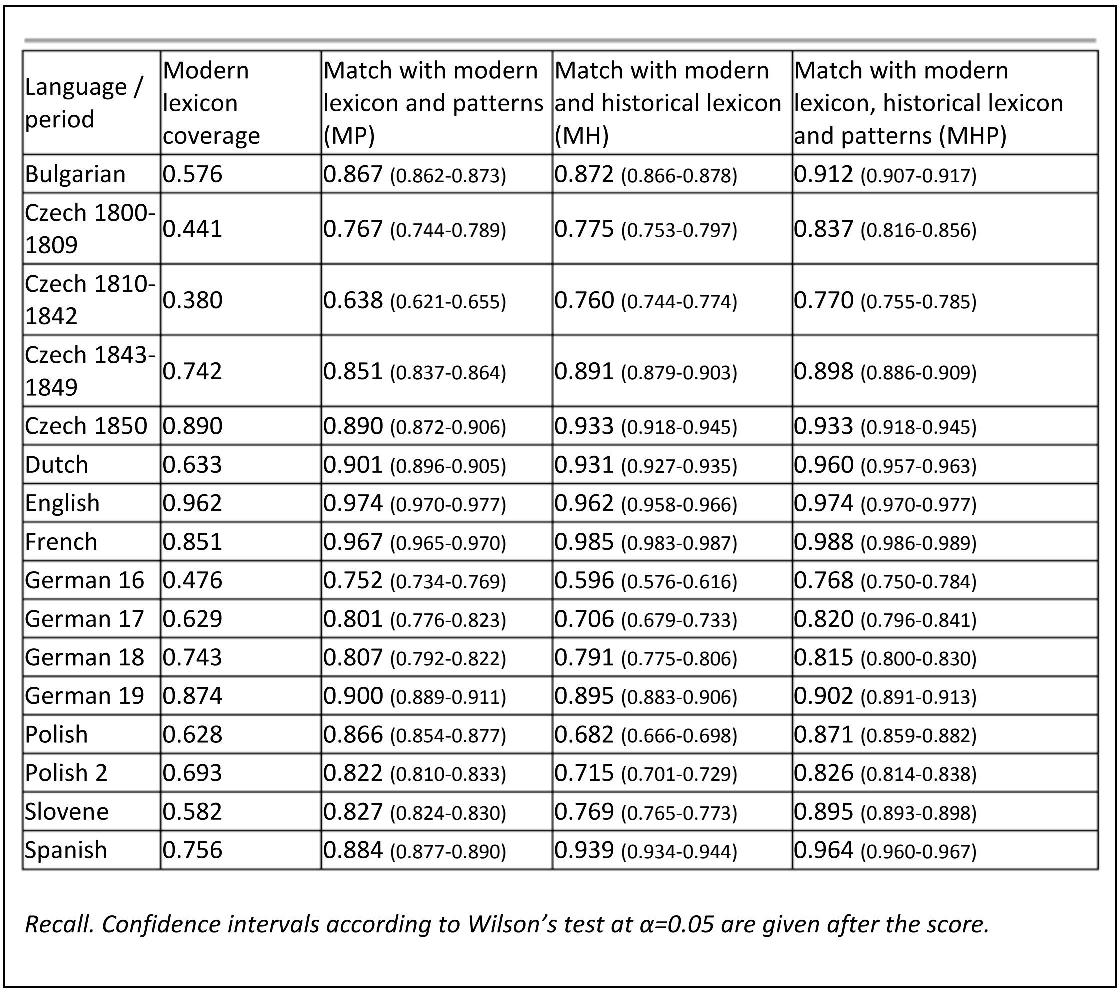
[49] http://www-psycho.uni-paderborn.de/lezius

[50] Cf. for instance http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-in-information-retrieval-1.html

form level is the more relevant, as the amount of ambiguity in the modern language has nothing to do with the degree to which we are able to bridge the "Historical Language Barrier". On the other hand, retrieval by lemma is maybe closer to a real-life IR scenario.

As it is, we are unable to perform modern word form assignment evaluation for all languages, since Czech, Dutch, English, Polish and Spanish did not annotate the word form equivalent in their IR evaluation datasets. On the other hand, the current German IR evaluation set does not annotate lemma for historical words that are identical to their modern equivalents. So for German, we evaluate modern word form assignment, and for all other languages, we evaluate modern lemma assignment.

6.2.3.1 Results



Language / period	Modern lexicon coverage	Match with modern lexicon and patterns (MP)	Match with modern and historical lexicon (MH)	Match with modern lexicon, historical lexicon and patterns (MHP)
Bulgarian	0.576	0.867 (0.862-0.873)	0.872 (0.866-0.878)	0.912 (0.907-0.917)
Czech 1800-1809	0.441	0.767 (0.744-0.789)	0.775 (0.753-0.797)	0.837 (0.816-0.856)
Czech 1810-1842	0.380	0.638 (0.621-0.655)	0.760 (0.744-0.774)	0.770 (0.755-0.785)
Czech 1843-1849	0.742	0.851 (0.837-0.864)	0.891 (0.879-0.903)	0.898 (0.886-0.909)
Czech 1850	0.890	0.890 (0.872-0.906)	0.933 (0.918-0.945)	0.933 (0.918-0.945)
Dutch	0.633	0.901 (0.896-0.905)	0.931 (0.927-0.935)	0.960 (0.957-0.963)
English	0.962	0.974 (0.970-0.977)	0.962 (0.958-0.966)	0.974 (0.970-0.977)
French	0.851	0.967 (0.965-0.970)	0.985 (0.983-0.987)	0.988 (0.986-0.989)
German 16	0.476	0.752 (0.734-0.769)	0.596 (0.576-0.616)	0.768 (0.750-0.784)
German 17	0.629	0.801 (0.776-0.823)	0.706 (0.679-0.733)	0.820 (0.796-0.841)
German 18	0.743	0.807 (0.792-0.822)	0.791 (0.775-0.806)	0.815 (0.800-0.830)
German 19	0.874	0.900 (0.889-0.911)	0.895 (0.883-0.906)	0.902 (0.891-0.913)
Polish	0.628	0.866 (0.854-0.877)	0.682 (0.666-0.698)	0.871 (0.859-0.882)
Polish 2	0.693	0.822 (0.810-0.833)	0.715 (0.701-0.729)	0.826 (0.814-0.838)
Slovene	0.582	0.827 (0.824-0.830)	0.769 (0.765-0.773)	0.895 (0.893-0.898)
Spanish	0.756	0.884 (0.877-0.890)	0.939 (0.934-0.944)	0.964 (0.960-0.967)

Recall. Confidence intervals according to Wilson's test at α=0.05 are given after the score.

The recall obtained using a witnessed historical lexicon in addition to spelling variations patterns and a modern lexicon (*HMP)*, shows a clear improvement both with respect to the baseline, and (in most cases) with respect to the modern-with-patterns approach (*MP*). In the case of Polish and nineteenth-century German, the difference between *MP* and *HMP* is small. This does not mean that IR lexicon building has been useless for these languages. The pattern sets for these languages were constructed and substantially refined during lexicon building, and are based on a large number of witnessed examples from the lexica. In a way, they could almost be regarded as a compressed representation of the lexicon.

Language / period	Match with modern lexicon and patterns	Match with modern and historical lexicon	Match with modern lexicon, historical lexicon and patterns
Bulgarian	0.551 (0.544-0.558)	0.751 (0.744-0.757)	0.562 (0.556-0.569)
Czech 1800-1809	0.792 (0.770-0.813)	0.773 (0.750-0.794)	0.778 (0.756-0.798)
Czech 1810-1842	0.814 (0.798-0.829)	0.799 (0.784-0.813)	0.791 (0.777-0.805)
Czech 1843-1849	0.817 (0.802-0.831)	0.812 (0.797-0.826)	0.810 (0.795-0.824)
Czech 1850	0.847 (0.827-0.865)	0.834 (0.814-0.852)	0.834 (0.814-0.852)
Dutch	0.427 (0.422-0.432)	0.448 (0.443-0.454)	0.344 (0.340-0.348)
English	0.351 (0.345-0.357)	0.471 (0.464-0.478)	0.351 (0.345-0.357)
French	0.751 (0.746-0.755)	0.731 (0.726-0.735)	0.723 (0.718-0.728)
German 16	0.705 (0.687-0.723)	0.892 (0.875-0.906)	0.702 (0.684-0.720)
German 17	0.783 (0.758-0.806)	0.936 (0.917-0.951)	0.779 (0.754-0.802)
German 18	0.806 (0.791-0.821)	0.972 (0.964-0.978)	0.805 (0.789-0.819)
German 19	0.878 (0.866-0.890)	0.985 (0.980-0.989)	0.877 (0.865-0.889)
Polish	0.565 (0.551-0.579)	0.476 (0.462-0.491)	0.497 (0.484-0.510)
Polish 2	0.597 (0.584-0.609)	0.529 (0.516-0.542)	0.541 (0.529-0.554)
Slovene	0.477 (0.474-0.481)	0.587 (0.584-0.591)	0.457 (0.454-0.460)
Spanish	0.684 (0.675-0.692)	0.706 (0.698-0.714)	0.627 (0.619-0.635)

(Unranked) precision. Confidence intervals according to Wilson's test at α=0.05 are given after the score.

The precision (and the F1) results are on the low side. This is mainly the result of the fact that, due to the ambiguity inherent in the lexica and the use of the patterns, we return a relatively long list of candidates. The precision scores fail to reflect that in many cases we are able to assign, with some success, a plausibility ranking to the candidates. We may consider an exact match in the historical lexicon more relevant than an exact match in the modern lexicon, and an exact match in the modern lexicon may be considered more plausible than a match using variation patterns[51]. Finally, different pattern-based matches may be ranked by weights of the relevant patterns.

A simple way to take the ranking into account[52] is to aggregate precision and recall in the *average reciprocal rank* of the first correct suggestion.

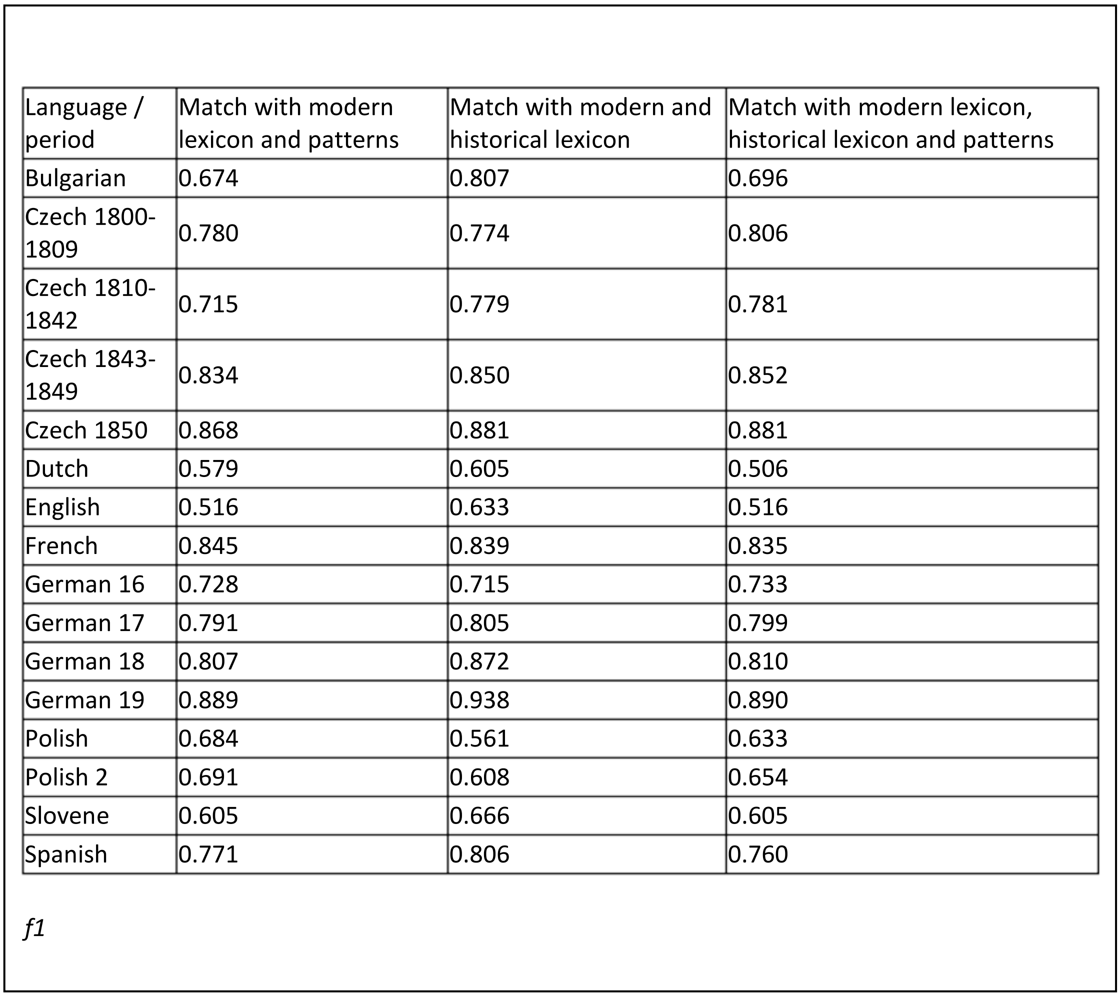
We define, for each token t:

RR(t):= 1/(rank of first correct suggestion) if *t* has a correct lemma suggestion, *0* otherwise,

and simply average this over all tokens to obtain the average reciprocal rank.

[51] This last preference is called the "modern wins" strategy in (Gotscharek et al. 2009, 2010).

[52] On the assumption that there is only one correct lemma assignment for each token



Language / period	Match with modern lexicon and patterns	Match with modern and historical lexicon	Match with modern lexicon, historical lexicon and patterns
Bulgarian	0.674	0.807	0.696
Czech 1800-1809	0.780	0.774	0.806
Czech 1810-1842	0.715	0.779	0.781
Czech 1843-1849	0.834	0.850	0.852
Czech 1850	0.868	0.881	0.881
Dutch	0.579	0.605	0.506
English	0.516	0.633	0.516
French	0.845	0.839	0.835
German 16	0.728	0.715	0.733
German 17	0.791	0.805	0.799
German 18	0.807	0.872	0.810
German 19	0.889	0.938	0.890
Polish	0.684	0.561	0.633
Polish 2	0.691	0.608	0.654
Slovene	0.605	0.666	0.605
Spanish	0.771	0.806	0.760

f1

Language / period	Match with modern lexicon and patterns	Match with modern and historical lexicon	Match with modern lexicon, historical lexicon and patterns
Bulgarian	0.766	0.804	0.839
Czech 1800-1809	0.724	0.735	0.794
Czech 1810-1842	0.602	0.734	0.744
Czech 1843-1849	0.806	0.862	0.868
Czech 1850+	0.849	0.902	0.902
Dutch	0.824	0.896	0.922
English	0.947	0.938	0.947
French	0.896	0.942	0.945
German 16	0.742	0.594	0.757
German 17	0.792	0.706	0.809
German 18	0.806	0.791	0.813
German 19	0.900	0.895	0.902
Polish	0.786	0.559	0.734
Polish 2	0.744	0.617	0.719
Slovene	0.743	0.758	0.874
Spanish	0.836	0.875	0.898

average reciprocal rank of first correct match

6.2.4 IR style evaluation of ranked retrieval by lemma

As remarked in the previous section, precision, recall and F1 score do not reflect the ranking of search results. One way of taking this into account is the so-called *Mean Average Precision*.

The average precision of a result with respect to a query *q* is defined by the average of the precision values measured at different recall levels. It can be easily computed as follows. Suppose *C* is the number of truly relevant items for *q,* and that the search engine returns items $d_1..d_m$ on *q.*

Now define the *precision at k, P_k* as

$$P_k = \frac{\text{number of relevant items among d1..dk}}{k}$$

Let $\delta(k) = 1/C$ if d_k is relevant, 0 otherwise. Then the *average precision* for *q* is

$$\sum_{k=1..m} \delta(k) * P_k$$

The *Mean Average Precision* of a retrieval system on a set *Q* of queries is simply the average of the average precision over all queries in *Q*. Note that besides taking ranking into account, this measure also differs from the ones in the previous section by averaging over distinct word *types* rather than tokens, which makes good coverage scores harder to achieve[53].

The table below contains measurements with *Q* = (set of true lemmata occurring in the corpus)

[53] And it also makes this score, on average, decreasing with corpus size if we take Q as defined in this section.

Language / period	Match with modern lexicon and patterns	Match with modern and historical lexicon	Match with modern lexicon, historical lexicon and patterns
Bulgarian	0.757	0.685	0.784
Czech 1800-1809	0.715	0.693	0.785
Czech 1810-1842	0.524	0.654	0.668
Czech 1843-1849	0.769	0.821	0.832
Czech 1850	0.879	0.921	0.921
Dutch	0.847	0.799	0.881
English	0.903	0.866	0.903
French	0.900	0.935	0.951
German 16	0.681	0.542	0.700
German 17	0.770	0.713	0.798
German 18	0.776	0.750	0.790
German 19	0.854	0.841	0.857
Polish	0.772	0.534	0.777
Polish 2	0.699	0.578	0.707
Slovene	0.536	0.366	0.581
Spanish	0.781	0.803	0.871

mean average precision

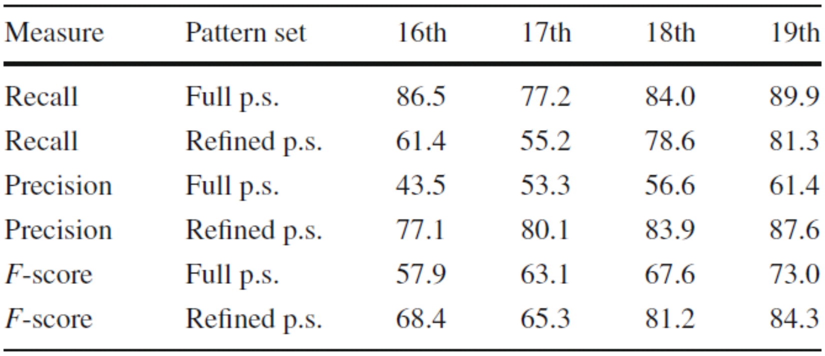
6.3 Possibilities for further development

The lexica built in IMPACT are partly "Proof of Concept" lexica, built with limited resources. Furthermore, the deployment tools can be refined to take, for instance, context information into account. One might wonder what procedure to take in developing and extending the lexica and the tools. Can the (limited) evaluation in this section point to the best ways to improve retrieval? We shall briefly discuss some possibilities, and clarify whether the evaluation data at hand can help to point us in the right direction. [onderstaande 5 punten komen deels maar lang niet helemaal overeen met de navolgende paragraaftitels, is verwarrend]

1. Making pattern matching more specific
2. Using (document or period-specific?) pattern weights
3. Corpus-based extension of witnessed historical lexica
4. Extending modern lexica
5. Taking sentence context into account

6.3.1 Making pattern matching more specific

This is discussed in (Gotscharek et al. 2010). At the cost of some loss of recall, it pays off in terms of precision to reduce the patterns to subsets attested in a certain period of time.



Measure	Pattern set	16th	17th	18th	19th
Recall	Full p.s.	86.5	77.2	84.0	89.9
Recall	Refined p.s.	61.4	55.2	78.6	81.3
Precision	Full p.s.	43.5	53.3	56.6	61.4
Precision	Refined p.s.	77.1	80.1	83.9	87.6
F-score	Full p.s.	57.9	63.1	67.6	73.0
F-score	Refined p.s.	68.4	65.3	81.2	84.3

Table (from Gotscharek et al. 2010, based on different data): recall, precision, and F-measure obtained when using the matching approach based on the full pattern set and based on a refined set of patterns adapted to the respective historical period.

6.3.2 Using (document or period-specific?) pattern weights

An alternative to 1 is to use frequency-based weights on patterns for the ranking of matching candidates. As an example, we can improve unranked precision for sixteenth-century German from 66 to 70% by pattern weighting and elimination of matches with a high combination of pattern weights, with no loss of recall.

6.3.3 Corpus-based extension of witnessed historical lexica

This is, of course, an option when background historical corpora are available. The option has been investigated for German in (Gotscharek et al. 2010). We quote their conclusions for Early Modern German:

> On the other hand, regardless of the corpus that is used it seems difficult to push recall beyond a certain limit: even with a lexicon of 50,000 words we only cover 50.66% of the missed vocabulary. We currently do not have a final explanation for this phenomenon. One reason is that the Additional Corpus is not large enough. In our work, we found that lexicon building with words that have a corpus frequency of three or less leads to negligible additional recall in most cases. This limit is reached with approximately 30,000 entries processed through the frequency-based approach. A second problem is the extremely high variance of the orthography due to missing standardization in the Early New High German period.

6.3.4 Extending modern lexica

An important weakness in our current evaluation is that we do not test the situation in which a lemma is unknown to both the modern and the historical lexicon. As can be seen in the table below, recall on the set of “known lemmata” is usually quite acceptable:

Language / period	Match with modern lexicon and patterns	Match with modern and historical lexicon	Match with modern lexicon, historical lexicon and patterns
Bulgarian	0.915	0.920	0.962
Czech 1800-1809	0.882	0.868	0.939
Czech 1810-1842	0.889	0.921	0.936
Czech 1843-1849	0.915	0.944	0.951
Czech 1850	0.944	0.962	0.962
Dutch	0.913	0.943	0.973
English	--	--	--
French	0.978	0.990	0.993
German 16	0.906	0.713	0.919
German 17	0.963	0.836	0.972
German 18	0.976	0.951	0.980
German 19	0.997	0.990	0.998
Polish	0.933	0.731	0.934
Polish 2	0.937	0.812	0.939
Slovene	0.919	0.823	0.963
Spanish	0.949	0.962	0.988

recall on items with lemma in modern lexicon

This shows that besides the witnessed historical lexicon, an extensive modern lexicon is also of great importance. Another way of improving recall could be to use a hypothetical modern lexicon obtained by reverse lemmatization of a relevant set of lemmata. In fact, applying reverse lemmatization to the user query would be a resource-free way of enhancing recall here.

6.3.5 Taking sentence context into account

We considered including part-of-speech tagging for historical language not feasible for IMPACT. One should bear in mind that all linguistic processing in IMPACT has to be relevant for both IR and OCR, and this is not the case with tagging. Nevertheless, modern to historical matching and part-of-speech tagging are tasks that are best combined. Within IMPACT, a tool chain (ToTrTaLe) combining these steps has been set up by JSI; cf. section 3.8.

7 Conclusions

After working for four years on German and Dutch, and for two years on a set of nine European Languages, we are in a position to evaluate the results of the work on lexicon building in IMPACT.

One conclusion is that the main vision and the tools foreseen in the original description of work were on the whole valid. This does not mean that there have been no bends and pitfalls along the road. All major tools have been adapted to a certain degree in the final years of the project.

1. Lexicon building from unclean OCR corpora turned out to be more important than we had foreseen. Lexicon building tools have been adapted to this.

2. Construction of OCR lexica from corpus data with frequency information was far less obvious than foreseen
3. It had not been foreseen that the ground truth production process would be as complex and time-consuming as it turned out to be. In the end, we can regard the workflow for ground truth production as a major IMPACT contribution to the community.
4. Different types of resource have been used by the IMPACT partners for lexicon building[54]; there is no unique “best approach” to lexicon development for OCR and IR of historical documents. There is for instance no clear winner when corpus-based lexicon building is compared to dictionary-based lexicon building. The following table summarizes which resources have been used by the partners.

8 References

Sayeed G. Choudhury, T. Dilauro, R. Ferguson, M. Droettboom, I. Fuginaga, “Document recognition for a million books”. In: *D-Lib Magazine*, Vol. 12, No. 3. (2006)

C. Clausner, S. Pletschacher, A. Antonacopoulos, Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments, *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)*, Beijing, China, September 2011, pp. 48-52.

Hugh Craig and R. Whipp (2010), Old spellings, new methods: automated procedures for indeterminate linguistic data, *Lit Linguist Computing* (2010) 25(1): 37-52 doi:10.1093/llc/fqp033

A. Ernst-Gerlach and N. Fuhr. “Generating search term variants for text collections with historic spellings.” In: *Proceedings of the 28th European Conference on Information Retrieval Research* (ECIR 2006). Springer, 2006

A. Ernst-Gerlach and N. Fuhr. “Retrieval in text collections with historic spelling using linguistic and spelling variants.” In *JCDL ’07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 333–341, New York, NY, USA, 2007. ACM

Erjavec, T., C. Ringlstetter, M. Žorga and A. Gotscharek, "Towards a Lexicon of XIXth Century Slovene". IS-JT '10 conference (14-15 October, Ljubljana, Slovenia)

Erjavec, T., C. Ringlstetter, M. Žorga, A. Gotscharek. A lexicon for processing archaic language: the case of XIXth century Slovene. *WoLeR 2011 at ESSLLI, International Workshop on Lexical Resources (*1-5 August, 2011, Ljubljana, Slovenia)

Erjavec, T. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities*, pages 33-38, Portland, OR, USA, 24 June 2011.

M. Federico, N. Bertoldi, M. Cettolo, *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, Proceedings of Interspeech*, Brisbane, Australia, 2008 (http://www.keithv.com/software/srilm/)

Lenz Furrer, Martin Volk, “Reducing OCR Errors in Gothic-Script Documents”. In: *Proceedings of the RANLP 2011 workshop on Language Technologies for Digital Humanities and Cultural Heritage* (2011), pp. 97-103.

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter and Klaus U. Schulz. Enabling Information Retrieval on Historical Document Collections - the Role of Matching Procedures and Special Lexica. *Proceedings of the ACM SIGIR 2009 Workshop on Analytics for Noisy Unstructured Text Data (AND 2009)*, Barcelona, 2009.

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter and Klaus U. Schulz. Constructing a Lexicon from a Historical Corpus. *Conference of the American Association for Corpus Linguistics (AACL09)*, Edmonton 2009.

[54] Cf the first table of section 4.9 for details

Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter and Klaus U. Schulz. On Lexical Resources for Digitization of Historical Documents. *The 9th ACM Symposium on Document Engineering (DOCENG 2009).*

Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter, Klaus U. Schulz, Andreas Neumann, "Towards information retrieval on historical document collections: the role of matching procedures and special lexica". In: *International Journal on Document Analysis and Recognition* 14, pp. 159-171. Springer 2010.

Claus Gravenhorst: *Applied IMPACT - Does the new FineReader Engine and Dutch lexicon increase OCR accuracy and production efficiency?* A case study by KB and CCS, Final IMPACT Conference, London, 2011-10-24

Guenthner, F., (1996): Electronic Lexica and Corpora Research at CIS, *International Journal of Corpus Linguistics*, 1(2), 1996.

van Halteren H., Rem M. (2009) . A tagger-lemmatiser for 14th century Dutch charters. Unpublished Paper Presented at Computational Linguistics in the Netherlands 2009. (*CLIN 2009*, Groningen, 22/01/2009).

S. Kempen, W. Luther, and T. Pilz. "Comparison of distance measures for historical spelling variants." In: *Artificial Intelligence in Theory and Practice*, volume 217 of IFIP International Federation for Information Processing, pp. 295–304. Springer Boston, 2006

Kestemont, M. Daelemans, W. & De Pauw G., 'Weigh your words - Memory-Based Lemmatization for Middle Dutch', in: *Literary and Linguistic Computing* 25:3 (2010), 287-30

M. Koolen, F. Adriaans, J. Kamps, and M. Rijke. "A cross-language approach to historic document retrieval." In M. L. et al., editor, *Proceedings of 28th European Conference on Information Retrieval Research (ECIR 2006)*, pp. 407–419. Springer, 2006.

Gary E. Kopec, Maya R. Said, Kris Popat, N-Gram Language Models for Document Image Decoding, Proc. SPIE 4670, 191 (2001).

Tomasz Lisowski. Pisownia polska. Główne fazy rozwoju (propozycja rozdziału podrecznika do nauczania tresci historycznojezykowych na studiach I stopnia) *Kwartalnik Jezykoznawczy* 2010/3-4 (3-4)

Maier-Meyer, P., (1995): *Lexikon und automatische Lemmatisierung*, PhD thesis, CIS, University of Munich.

Neudecker, C., S. Schlarb, Z. M. Dogan, P. Missier, S. Sufi, A. Williams and K. Wolstencroft. An experimental workflow development platform for historical document digitisation and analysis. In: *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing* (HIP '11). ACM, New York, NY, USA, 161-168. DOI=10.1145/2037342.2037370

Thomas Pilz, Andrea Ernst-Gerlach, Sebastian Kempken, Paul Rayson, and Dawn Archer (2008) The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic? *Lit Linguist Computing* (2008) 23(1): 65-72 doi:10.1093/llc/fqm044

Ray Smith, Limits on the application of frequency-based language models to OCR, 2011 International Conference on Document Analysis and Recognition

Luc Vincent (2007), Google Book Search: Document Understanding on a Massive Scale, *Proc. International Conference on Document Analysis and Recognition, ICDAR'2007*, Curitiba, Brazil, Sep. 2007, http://www.vincent-net.com/luc/papers/07icdar_googlebooks.pdf

9 Appendix: Evaluation data used in this report

9.1 Bulgarian

1. Bylgarska iliustracia, 1880 -
2. Jenski glas, 1889 -
3. Sborniche za spomen na 25-godishninata ot smyrtta na Levski, 1898
4. Spisanie Dennica, 1890
5. Ugozapadna Bulgaria, 1893 -
6. Zelokupna Bulgaria, 1880 -

9.2 Czech

1. Co jest konstituce?, čili, Krátký, prostonárodní wýklad hlawnějších zásad konstitucí ewropejských, 1848
2. Ferina Lišák z Kuliferdy a na Klukově, čili, Kratičká historye zlopověstných kousků starého Reinecke, 1848
3. Homerowa Iliada, 1802
4. Na den narození neimocněišího, a neijasněišího cysare rímského, téz dědičného rakauského a krále ceského, Frantiska II., w Praze 12. den mesyce Unora, léta 1805, 1805
5. Plody sborů učenců řeči ceskoslowanské prešporského, 1836
6. Rozprawy o gmenách, počátkách i starožitnostech národu Slawského a geho kmeni /, 1830
7. Sokol, 1872
8. Základowé pitwy (Anatomie), čili, Soustawnj rozbor a popis těla lidského a gednotliwých geho částek, 1840

9.3 Dutch

The evaluation subset of the IMPACT dutch demonstrator dataset includes the following titles which have been included in the OCR evaluation:

1. 's Hertogenbossche courant, 1771
2. Affiches, annonces et avis divers d'Amsterdam = Advertentiën, aankondigingen en verschillende berigten van Amsterdam, 1813
3. Algemeen Handelsblad, 1866
4. Arnhemsche courant, 1817
5. Dagblad van 's Gravenhage, 1827
6. Dagblad van Zuidholland en 's Gravenhage, 1858-1858
7. De Nederlander: nieuwe Utrechtsche courant: (staatkundig- nieuws-, handels- en advertentie-blad) / onder red. van J. van Hall, 1853-1853
8. Extra tyding. Extract uit de resolutien der heeren Staaten van Holland [...] genomen op [...] 4. september 1786, 1786-1786
9. Feest en lydens stoffen voor de hervormde gemeente te Alkmaar, 1787-1787
10. Handelingen der Staten-Generaal 1826-1827, 1826-1827
11. Handelingen der Staten-Generaal 1857-1858, 1857-1858
12. Handelingen der Staten-Generaal 1868-1869, 1868-1869
13. Handelingen der Staten-Generaal 1891-1892, 1891-1892
14. Handelingen der Staten-Generaal 1929-1930, 1929-1930
15. Handelingen der Staten-Generaal 1931-1932, 1931-1932
16. Handelingen der Staten-Generaal 1932-1933, 1932-1933
17. Handelingen der Staten-Generaal 1933, 1933-1933
18. Handelingen der Staten-Generaal 1934-1935, 1934-1935

19. Handelingen der Staten-Generaal 1935-1936, 1935-1936
20. Handelingen der Staten-Generaal 1936-1937, 1936-1937
21. Handelingen der Staten-Generaal 1937-1938, 1937-1938
22. Handelingen der Staten-Generaal 1938-1939, 1938-1939
23. Het nieuws van den dag: kleine courant, 1870-1870
24. Journal de la province de Limbourg, 1824-1824
25. Kort begrip der waereld-historie voor de jeugd. / By J.F. Martinet, 1789-1789
26. Middelburgsche courant, 1774-1774
27. Nederlandsche staatscourant, 1827-1827
28. Nieuwe Rotterdamsche courant: staats-, handels-, nieuws- en advertentieblad, 1854-1854
29. Oprechte Haarlemse courant, 1805-1805
30. Provinciale Overijsselsche en Zwolsche courant: staats-, handels-, nieuws- en advertentieblad, 1852-1852
31. Rechtsgeleerd advis in de zaak van den gewezen stadhouder, en over deszelfs schryven aan de gouverneurs van de Oost- en West-Indische bezittingen van den staat [...]. Ingelevert [...] op den 7 january 1796. / By B. Voorda et al, 1796-1796
32. Verhaal van het levensgevaar, waar in zig drie Rotterdamsche burgers [...] bevonden hebben, te Utrecht, 1784-1784
33. Vrijmoedige aanmerkingen, over de uitsluiting van allen die door publieke armkassen bedeeld worden, als stemgerechtigden [...] bij eene oproeping van het Nederlandsche volk tot eene Nationaale Conventie, 1795-1795

9.4 French

1. "Trois discours philosophiques: le I, de la comparaison de l'homme avec le monde ; le II, du principe de la génération de l'homme ; le III, de l'humeur mélancolique, mis de nouveau en lumière par Jourdain Guibelet", 1603
2. Discours de la méthode... plus la dioptrique, les météores, la méchanique et la musique, qui sont des essais de cette méthode, par René Descartes. Avec des remarques et des éclaircissements nécessaires, 1668
3. Dissertation de la philosophie en général, 1668
4. Le Prince instruit en la philosophie, en françois... avec une métaphysique historique... par messire Besian Arroy,... Première édition, 1671
5. Les Méditations métaphysiques de René Descartes touchant la première philosophie. 2e édition reveüe et corrigée par le traducteur (Charles d'Albert, duc de Luynes) et augmentée de la version d'une lettre de M. Des Cartes au R. P. Dinet..., 1661
6. Lettre de M. Gadroys à M. de La Grange Trianon,... pour servir de réponse à celle que M. de Castelet a écrite contre les raisons de M. Descartes touchant le flux et le reflux de la mer. - Seconde lettre de M. Gadroys... [au même, sur le même sujet.], 1677
7. Traitté de l'esprit de l'homme, de ses facultez et fonctions, et de son union avec le corps, suivant les Principes de René Descartes, par Louis de La Forge,..., 1661

9.5 English

1. A speech, or complaint, lately made by the Spanish embassadour to his Majestie at Oxford, upon occasion of the taking of a ship called Sancta Clara in the port of Sancto Domingo, richly laden with plate, cocheneal and other commodities of great value, by, 1643
2. A treatise touching falling from grace. Or Thirteen arguments tending to prove that believers cannot fall from grace, as they were laid down at a conference at Yalding in Kent, examined and answered, with many absurdities of that doctrine shewed. Whereunt, 1653

3. Aberdeen Journal, 1821
4. Calendar-reformation. Or, An humble addresse to the Right Honorable the Lords and Commons assembled in Parliament, touching the dayes and moneths, that they may be taught to speak such a language as may become the mouth of a Christian. / By I.B., 1648
5. Cases and questions resolved in the civil-lavv. Collected by R. Zouch professor of the civil-law in Oxford., 1652
6. Certain information from Devon and Dorset: concerning the Commission of Array., 1642
7. Exeter Flying Post, 1894
8. Miscellany in Verse and Prose, 1739
9. Northern Echo, 1866
10. Old Poor Robin … an almanac, 1777
11. Paidon nosemata· = or Childrens diseases both outward and inward. From the time of their birth to fourteen years of age. With their natures, causes, signs, presages and cures. In three books: 1. Of external 2. Universal 3. Inward diseases. Also, the resol, 1664
12. Playbills, 1850-ish
13. Prologus Here begynneth the prologue of the storye of Thebes, 1497
14. Reflections on the different ideas of the French and English in regard to cruelty ... By a man [H. W.], 1759
15. The Country Journal or The Craftsman, 1742
16. The Examiner, 1808
17. The Hull Packet & Humber Mercury, 1828
18. The Maidstone Garland, 1740
19. The Morning Chronicle, 1843

The seventeenth century subset uses following titles:

1. "Areopagitica; a speech of Mr. John Milton for the liberty of vnlicens'd printing, to the Parlament of England", 1644
2. "VVil: Bagnal's ghost. Or the merry devill of Gadmunton. In his perambulation of the prisons of London. / By E. Gayton, Esq;.", 1655
3. A speech, or complaint, lately made by the Spanish embassadour to his Majestie at Oxford, upon occasion of the taking of a ship called Sancta Clara in the port of Sancto Domingo, richly laden with plate, cocheneal and other commodities of great value, by, 1643
4. A treatise touching falling from grace. Or Thirteen arguments tending to prove that believers cannot fall from grace, as they were laid down at a conference at Yalding in Kent, examined and answered, with many absurdities of that doctrine shewed. Whereunt, 1653
5. Calendar-reformation. Or, An humble addresse to the Right Honorable the Lords and Commons assembled in Parliament, touching the dayes and moneths, that they may be taught to speak such a language as may become the mouth of a Christian. / By I.B., 1648
6. Cases and questions resolved in the civil-lavv. Collected by R. Zouch professor of the civil-law in Oxford., 1652
7. Certain information from Devon and Dorset: concerning the Commission of Array., 1642
8. Hollands ingratitude, or, A serious expostulation with the Dutch shewing their ingratitude to this nation, and their inevitable ruine, without a speedy compliance and submission to His Sacred Majesty of Britain / by Charles Molloy of Lincolns-Inn, Gent., 1666
9. Paidon nosemata· = or Childrens diseases both outward and inward. From the time of their birth to fourteen years of age. With their natures, causes, signs, presages and cures. In three books: 1. Of external 2. Universal 3. Inward diseases. Also, the resol, 1664
10. The golden trade: or, A discouery of the riuer Gambra, and the golden trade of the Aethiopians Also, the commerce with a great blacke merchant, called Buckor Sano, and his report of the houses couered with gold, and other strange obseruations for the good, 1623

9.6 German

1. Das Buch des heyligen Römischen Reichs unnderhalltunge, 1501
2. Die Poesie ihr Wesen und ihre Formen mit Grundzügen der vergleichenden Literaturgeschichte, 1884
3. Echo Deß Hochzeitlichen Te Deum Laudamus, 1722
4. Ergebnisse der Erhebungen über die Beschäftigung gewerblicher Arbeiter an Sonn- und Festtagen, Bd.:1, Gruppe I bis VII der Gewerbestatistik, Berlin, 1887, 1887
5. Quedlinburgisches Kreis-Tags-Memorial, 1673
6. Von der Regierung der Kirche und den unterschiedlichen Würden der Geistlichkeit *(full title in comments), 1779
7. Warhaffter und grundlicher Bericht uß was Ursachen Martinus du Voysin (zu Basel verburgerter Krämer) inn der Statt Surseew im Aargöw, ..., den 13. Tag Octobris deß 1608. Jars erstlich enthauptet, und volgends verbrennt worden, 1609

9.7 Polish

1. Adwersaria, albo terminata sprawy wojennej, która się toczyła w wołoskiej ziemi z tureckim cesarzem, 1621
2. Chorągiew Sarmacka w Wołoszech, to jest pospolite ruszenie i szczęśliwy powrót Polaków z Wołoch w roku 1621, 1621
3. Diariusz wiadomości od wyjazdu króla z Wilna do Smoleńska, 1610
4. Discurs o cenie pieniedzy teraznieyszey y o niektorych skutkach iey..., 1632
5. Nowe Ateny, albo Akademia wszelkiey scyencyi pełna, na różne tytuły iak na classes podzielona, mądrym dla memoryału, idiotom dla nauki, politykom dla praktyki, melancholikom dla rozrywki erygowana ... . Część 3 albo Supplement., 1746
6. Pasja żołnierzy obojga narodów w stolicy moskiewskiej krótko opisana, 1613
7. Powodzenia niebezpiecznego ale szczęśliwego wojska j. k. m. w Multanach opisanie, 1601
8. Relacja chwalebnej ekspedycji Jana Kazimierza, króla polskiego i szwedzkiego, 1650
9. Wyprawa i wyjazd sułtana Amurata, cesarza tureckiego, na wojnę do Korony Polskiej, 1634
10. Wyprawa i wyjazd sułtana Amurata, cesarza tureckiego, na wojnę do Korony Polskiej_(bitonal version), 1634
11. Żałosne opisanie upadku króla hiszpańskiego na morzu i na lądzie, 1589

9.8 Slovene

1. Genovefa, 1841
2. Gosp. Krištofa Šmida korarja avgustanskiga, zgodBe S. Pisma za mlade ljud..., 1850
3. Kmetijske in rokodelske novice, 1844
4. Kratkozhasne uganke, 1788
5. Kuharske Bukve, 1799
6. Marianske Kempensar, ali Dvoje bukuvze, 1769
7. Novice kmetijskih, rokodelnih in narodskih reči, 1851
8. Sgodbe svetiga pisma za mlade ljudi, 1830
9. Ta male katechismus, 1768
10. Vezhna pratika od gospodarstva, 1789
11. Zerkviza na skali, 1855

9.9 Spanish

We split the corpus into three subsets, Development, Evaluation and Demonstration, as shown in the following table:

DEVELOPMENT

Author	Title	Pages	Prima ID
Juan Boscán	Las obras de Juan Boscán y algunas de Gracilasso de la Vega	500	004484946-00485445
San Juan de la Cruz	Obras de San Juan de la Cruz	971	00442916-00443886
Mateo Alemán	Guzmán de Alfarache	411	00439017-00439427
Luis de Góngora	Polifemo comentado	274	00441307-00441580
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […] Tomo primero. Que contiene las letras A.B.	826	00444481-00445306
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […] Tomo segundo. Que contiene las letras C.	725	00445307-00446031
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […] Tomo tercero. Que contiene las letras D.E.F	827	00446032-00446857 / 00484945
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […] Tomo quarto. Que contiene las letras G.H.I.J.K.L.M.N	707	00446858-00447564
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […] Tomo quinto. Que contiene las letras O.P.Q.R	667	00447565-00448231
Real Academia Española de la Lengua	Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad, con las phrases o modos de hablar […]Tomo sexto. Que contiene las letras S.T.V.X.Y.Z	613	00448239-00448851
	EVALUATION		
Garcilaso de la Vega	Obras de Garcilaso con anotaciones de Francisco Sánchez Brocense	285	00438732-00439016
Lope de Vega	Obras de Lope de Vega	759	00485446-00486204

Sor Juana Inés de la Cruz	Carta Atenagórica	35	00443887-0044392
Inca Garcilaso de la Vega	Comentarios reales	546	00439428-00439973
Anónimo	Lazarillo de Tormes	140	00440435-00440574
Francisco de Quevedo	Parnasso español	671	00441581-00442251
	DEMONSTRATION		
Santa Teresa de Jesús	Obras de Santa Teresa	559	00443922-00444480
Jorge Juan	Observaciones Astronómicas y Físicas	461	00439974-00440434
Miguel de Cervantes	El Quijote	664	00442252-00442915
Pedro Calderón de la Barca	Obras de Calderón de la Barca	486	00437807-00438292

Also, these books present several defects and enhancement characteristics, some of the most frequent being bad printing, narrow binding, shining through, stamps and stains or uncropped and skewed images.