Maria Sukhareva, Zahurul Islam, Armin Hoenen, Alexander Mehler

A Three-step Model of Language Detection in Multilingual Ancient Texts

Ancient corpora contain various multilingual patterns. This imposes numerous problems on their manual annotation and automatic processing. We introduce a lexicon building system, called Lexicon Expander, that has an integrated language detection module, Language Detection (LD) Toolkit. The Lexicon Expander post-processes the output of the LD Toolkit which leads to the improvement of f-score and accuracy values. Furthermore, the functionality of the Lexicon Expander also includes manual editing of lexical entries and automatic morphological expansion by means of a morphological grammar.

1 Introduction

For more than a decade, ancient languages have been an object of research in computational humanities and related disciplines (Smith et al., 2000; Bamman et al., 2008; Bamman and Crane, 2009; Gippert, 2010a). This relates to building morphosyntactic resources (Passarotti, 2000; Koster, 2005), co-occurrence networks (Büchler et al., 2008; Mehler et al., 2011a) and dependency treebanks, which often focus on texts in Latin or Greek (Bamman and Crane, 2009; Passarotti, 2010). As these efforts concern dead and, thus, low-resource languages, the morphological, syntactic and semantic analysis of them is a challenging task. As a consequence, *manual* annotation is an indispensable companion of building resources out of ancient texts that can be used as reliable input to the various tasks of NLP. This holds especially for ancient languages as, for example, Avestan, Old High German (OHG) or Old Georgian (Gippert, 2006), which – unlike Latin and Greek – are less common objects of computing in the humanities. In these cases, corpus annotation is often accompanied by the manual generation of a full-form lexicon as a prerequisite of building lemmatizers and taggers for these languages.

A central challenge of annotating such corpora together with building corpus-specific lexica is the multilingualism of ancient texts. This relates to texts that contain word forms of different languages as a result of, for example, ancient annotations (Migne, 1855) or of fragments of different translations (Gippert, 2010b). An example of a corpus that mixes source texts with notes of different languages is the *Patrologia Latina (PL)* (Section 3). Another example are documents (e.g., in OHG) that contain a multitude of words borrowed from another language (e.g., Latin). In all these cases, corpus building and lexicon formation are faced with the task of detecting and separating the corresponding source languages correctly – *starting from the level of tokens via the level of sentences up to the level of whole paragraphs.* This is needed since any subsequent

JLCL 2011 – Band 26 (2) – 167-179

step of preprocessing (e.g., lemmatization or PoS tagging) is sensitive to the underlying language. Thus, in order to apply different preprocessors in a language-specific manner, one needs to know the language of any segment of the input texts. Moreover, the build-up of full-form lexica is by and large a task of manual annotation since taggers are hardly available for low-resource languages such as Old Georgian or OHG. In these cases, one needs to prevent any human annotator from handling, for example, thousands of *French* word forms in a corpus such as the Patrologia Latina if the target language is *Latin*.

In this paper, we introduce a software system called *Lexicon Expander* that supports the build-up of full-form lexica for ancient languages. The Lexicon Expander is part of the *eHumanities Desktop* (Mehler et al., 2011b) (Gleim et al., 2009a) that has been built as an online system of corpus processing in the digital humanities. The Lexicon Expander provides an online interface for lemmatizing and expanding unknown words morphologically. A central ingredient of the Lexicon Expander is a three-step language detector that annotates each input word by the name of the language that it probably manifests. Using these annotations, any human annotator can select language specific subsets of unknown words to handle them separately while leaving behind all words that do not belong to the target language of the lexicon to be built. We evaluate this model in two experiments: the one being based on ancient corpus data, the other being based on samples of modern languages.

The paper is organized as follows, Section 2 briefly discusses other projects that deal with building historical lexica. Section 3 gives an overview of multilingualism in ancient corpora and problems connected with it. Section 4 describes corpus preprocessing. It introduces the Language Detection Toolkit and the Lexicon Expander. Section 5 provides evaluation results. Section 6 discusses findings and draws a conclusion.

2 Related Work

This section gives an overview of systems developed for creating and processing of historical lexica. An extensive work on building historical lexica was done within the framework of the *Improving Access to Text (IMPACT)* project (Balk, 2010). The aim of the project is to digitalize printed documents created before the 19^{th} century. Amongst others, the IMPACT project provides tools for named-entity recognition and lexicon building. As the main task of this project is to heighten the accessibility to historical corpora and to simplify the search process, they provide a toolbox that assigns modern lemmata to historical word forms in order to avoid search problems, caused by historical spelling variations and changes in inflectional morphology. Presently, they provide historical lexica for Dutch and German and a graphical user interface for named entity attestation.

ToTrTaLe (Tokenisation, Transcription, Tagging and Lemmatization) (Erjavec, 2011) is a tool, developed for automatic linguistic annotation and applied to 19^{th} century Slovene. The input of the tool is a TEI encoded corpus. The tool automatically does a morphosyntactic annotation, assigns lemmata and modern equivalents of the

words. Lemmatization is also done by means of assigning modern lemmata to historical word forms. The output of the system is a TEI document containing the annotation information. The historical word forms were manually verified (Erjavec et al., 2011). A specialized editor, called LexTractor (Gotscharek et al., 2011), was used for processing of the historical corpus of Slovene. This web-tool, introduced by (Gotscharek et al., 2009), builds historical lexica with word forms mapped to modern lemmata. The GUI allows to work with unknown words, found in the corpus, and to manually annotate them. A user is asked to accept or to reject readings, proposed by the system. The tool was applied to build a historical German corpus, which currently contains ca. 10,000 entries.

Unlike the lexicon builder, which is presented in this paper, none of the aforementioned tools enables a user to annotate the unknown words with their language. We are not aware of any lexicon building tool, which also applies language detection to historical corpora. In this sense, we provide a tool to "fill this gap".

3 Multilingualism in Ancient Texts

There are various sources of multilingualism in ancient corpora, starting with mere borrowings and ending with comments in foreign languages added by corpus editors. This section briefly discusses the challenges, which multilingualism imposes on the annotation of ancient corpora.

Patrologia Latina (PL) is a collection of documents dating from the 4th till the 13th century. The Patrologia Latina was published in the 19th century. The original printing plates were destroyed in 1868, but lately restored and new editions were published. The Patrologia Latina is comprised of 8,508 documents written by 2,004 authors. The corpus includes over 100 Mio tokens (Mehler et al., 2011a).

The PL Corpus was lemmatized, tokenized and tagged with parts-of-speech (Mehler et al., 2011b). Nevertheless, there are ca. 700 000 tokens in the PL corpus that are marked as unknown. In this case, all the unknown tokens are to be annotated manually, but a more precise look at the unrecognized tokens reveals that a great number of them are not Latin words. The reason of it is that we deal with editions of the Patrologia Latina.

(1) Reliquiae antiquae Scraps from ancient manuscripts illustrating chiefly early english litterature and the english language, edited by Thomas Wright Esq. London, in 8, 2 vol. Remling, Urkundliche Geschichte der ehemaligen Abteien und Klöster in Rheinbaiern Neustadt a.d. Haardt, 1838, in 8,1 - 2. Reschius, Annales ecclesiae Sabinensis. Augustae Vindelic., 1765 in folio, 1-3.

These editions include frequent editors' comments in French, English, German, Italian, Portuguese etc. (1) is an example taken from the PL Corpus, containing editors' comments in English and German and the name of a document in Latin. Filtering out all the unknown words of the foreign origin saves a great amount of annotator's efforts.

The multilingualism problem is found in other corpora as well. The OHG corpus, created in the framework of the TITUS project[1] (Gippert, 2001), is composed of 101 texts with over 400,000 tokens. Some OHG texts are direct translations of Latin texts. Therefore, Latin words, phrases and sentences occur often in such texts. All in all, we found that approximately 9% of the words in the OHG corpus are Latin and 67% of them are single words within a context of OHG. (2) is an example of a bilingual sentence found in the OHG corpus.

(2) **Ein relatio ist patris ad filium ánderiû ist filii ad**
One.OHG relation.L is.OHG father.L to.L son.L other.OHG is.OHG son.L to.L
patrem[. . .]
father.L[. . .]
"One relation is between the Father and the Son, the other one - is between the Son and the Father[. . .]"

We find similar examples in the Avestan corpus (Example (3)). The Avestan corpus features ca. 30,000 words. The texts are written in Avestan, but some text segments are directly translated into one of the following languages: Middle Persian, New Persian, Gujarati, Sanskrit and included in some documents as comments, translations or instructions.

(3) **az xvarəθəm miiazdəm tā ānōh 2 bār guft[. . .]**
from.MP food.AV food sacrifice.AV till.MP there.MP two.MP time.MP speak.MP
"from "food sacrifice" up to here it is to be said twice..."

Summing up, multilingualism in ancient texts is a phenomenon found on lexical, phrasal, sentential and textual levels. In other words, the language can vary from verse to verse or sentence to sentence. Single foreign words and phrases also occur in texts. Filtering out foreign words that do not belong to the target language would simplify numerous tasks of automatic and manual corpus processing, such as lexicon building, corpus annotation, collocation extraction etc.

4 Approach

This section introduces a system for language detection, called *Language Detection (LD) Toolkit*, and the Lexicon Expander, an application module for the eHumanities Desktop (Gleim et al., 2009b). Section 4.2 describes the integration of the LD Toolkit into the Lexicon Expander and the system architecture. Corpus preprocessing is described in Section 4.1. The LD Toolkit is presented in Section 4.3. Finally, we describe the Lexicon Expander, which processes the output of the LD Toolkit in 4.4.

4.1 Preprocessing

The input corpora are annotated with the help of the *PrePro2010 - Text Processing Tool*[1], a text preprocessing tool that automatically does lemmatization, tokenization,

[1] Thesaurus of Indo-European Text and Language Materials (TITUS), http://titus.uni-frankfurt.de.
[1] PrePro2010-Text Processing Tool:http://api.hucompute.org/preprocessor/

sentence boundary detection, stemming, parts-of-speech tagging and named entity recognition (Waltinger, 2010). The resulting TEI P5 files are uploaded in a repository on the *eHumanities Desktop*.

4.2 System Design

The *Lexicon Expander* is an application module, implemented in the framework of the *eHumanities Desktop*, which enables a user to create, organize and annotate lexica.

Figure 1 shows the architecture of the system. A multilingual ancient corpus is converted in a TEI P5 (Sperberg-McQueen and Burnard, 2009) format and saved in a repository in the eHumanitiesDesktop. The user chooses the preprocessed corpus from the repository (Figure 2b). The TEI P5 corpus contains information about lemmata and PoS. Words that were not analysed during the preprocessing are marked with the attribute *"function = unk"*. The Lexicon Expander processes the TEI P5 marked up corpus and extracts such unknown words.

When the unknown words are extracted, the system builds up the lexicon out of them. In order to simplify manual annotation by filtering out words, that do not belong to the target language, the language of the words can be detected before building up the lexicon. The language detection is implemented by means of the LD Toolkit 4.3. The LD Toolkit runs as a background process and the Lexicon Expander can send various input to it. Previous studies (Islam et al., 2011) showed that the LD Toolkit has low *f-score* and *accuracy* if it takes a single word as an input, but if the toolkit is applied to sentences, *f-score* and *accuracy* are high.

In terms of the language detection, the user can choose from three options. First of all, the user can decide not to apply any language detection. In this case lexical entries are added directly into the lexicon without any language assigned. The lexicon is saved into a MySQL database. The second option presupposes that the user can choose to apply the LD Toolkit without the Lexicon Expander model. Then the LD Toolkit gets unknown words as an input. The output of the LD Toolkit is directly saved into the MySQL database. After the GUI of the Lexicon Expander (Figure 2a) is refreshed, the language column in the Lexicon Expander will be filled. The last option is to apply the Lexicon Expander model. In this case, the input of the LD Toolkit is unknown words and sentences, in which these words occur. Finally the output of the LD Toolkit is post-processed by the Lexicon Expander as described in Section 4.4. The language of the unknown words is re-assigned and saved into the MySQL Database.

When lexical entries are saved, the GUI of the Lexicon Expander (Figure 2a) is refreshed and the lexicon is available for further editing (Figure 2c). It is possible to edit the language of a lexical entry manually, specify its part of speech and assign values to grammatical categories. The user can also apply morphological expansion by means of a morphological grammar, defined by a finite-state compiler FOMA (Hulden, 2009). FOMA can assign parts of speech and respective grammatical features. Once the user finished editing the lexical entry, it is saved into the lexicon.

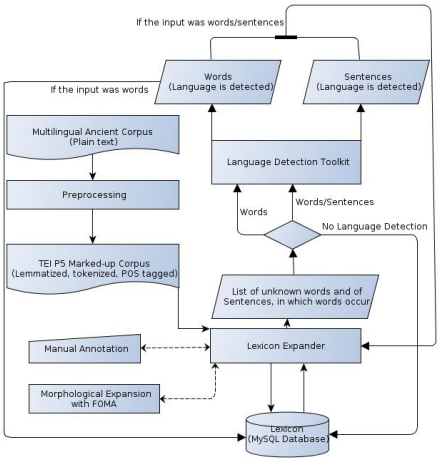


Figure 1: System diagram of the Lexicon Expander

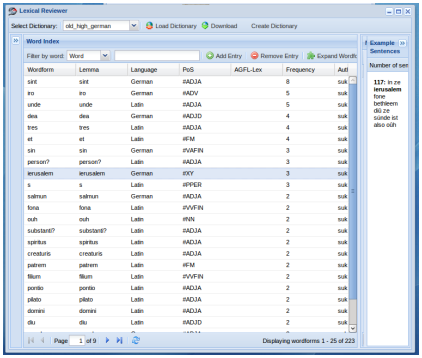
4.3 Language Detection Toolkit

The Language Detection (LD) Toolkit is based on Cavnar and Trenkle (1994) and Waltinger and Mehler (2009). Recently, Islam et al. (2011) have applied this technique to ancient languages. For every target category Islam et al. (2011) learn an ordered list of most frequent n-grams in descending order. The same is done for any input text stream so that categorization occurs by measuring the distance between n-gram profiles of the target categories and n-gram profiles of the input data. The idea behind this approach is that similar texts share features that are equally ordered.

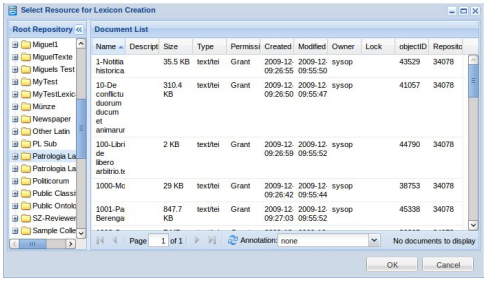
More specifically, classification is done by using the range of corpus features listed in (Waltinger and Mehler, 2009). Predefined information is extracted from the corpus to build sub-models based on these features. Each sub-model consists of a ranked frequency distribution of subsets of corpus features. The corresponding n-gram information is extracted for $n = 1$ to 5. Each n-gram gets its own frequency counter. The normalized frequency distribution of relevant features is calculated according to:

$$\widehat{f_{ij}} = \frac{f_{ij}}{\max_{a_k \in L(D_j)} f_{kj}} \in [0, 1] \tag{1}$$

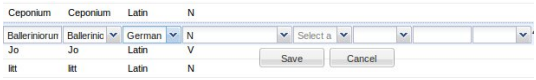
$\widehat{f_{ij}}$ is defined as the frequency f_{ij} of feature a_i in D_j divided by the frequency of the most frequent feature a_k in the feature representation $L(D_j)$ of document D_j (Waltinger and Mehler, 2009). To categorize any document D_m, it is compared to each category C_n using the distance d of the rank r_{mk} of feature a_k in the sub-model of D_m with the corresponding rank of that feature in the representation of C_n:



(a) Lexicon Expander



(b) Repository window



(c) Manual editing of lexical entries

Figure 2: The GUI of the Lexicon Expander

$$d\left(D_m, C_n, a_k\right) = \begin{cases} |r_{mk} - r_{nk}| & a_k \in L(D_m) \wedge a_k \in L(C_n) \\ max & a_k \notin L(D_m) \vee a_k \notin L(C_n) \end{cases} \quad (2)$$

$d\left(D_m, C_n, a_k\right)$ equals max if feature a_k does not belong to the representation of D_m or to the one of category C_n. max is the maximum that the term $|r_{mk} - r_{nk}|$ can assume.

To detect the language of a document, the LD toolkit traverses the document sentence by sentence and detects the language of each sentence (Islam et al., 2011). If the document is homogeneous, (i.e., all sentences belong to the same language), then sentence level detection suffices to trigger other processing tools (e.g., Parsing, Tagging and Morpho-syntactic analysis) that require language detection. In the case that the sentences belong to more than one language (i.e., in the case of a heterogeneous document), the toolkit processes the document word by word and detects the language of each token separately. This step is necessary in the case of multilingual documents that contain words from different languages even within the same sentences. For example: in a scenario of lemmatization or morphological analysis of a multilingual document, it is necessary to trigger language specific tools to avoid errors. Just one tool needs to be triggered for further processing of a homogeneous document, whereas for a heterogeneous document the same kind of tool has to be triggered based on the word level.

The LD toolkit is used as a web service component of the *Lexicon Expander* framework. Currently, the toolkit is able to detect 70 languages. It calculates the distances between

the input text and each of these 70 models. It returns the model name that minimizes distance. The input can be a document, a sentence or a word. The user can restrict the number of models to be used to detect an input. For example, in the case of a multilingual document, which contains sentences from OHG, Latin and Old Saxon languages, the toolkit can be restricted to these three models. In the case of ambiguity, the LD toolkit returns a list of languages with scores that can be used by the *Lexicon Expander* for further processing to assign the correct target language of the input.

4.4 The Lexicon Expander

The LD toolkit can be applied to a word, a sentence or a whole document. It yields high classification results, when its input is a document or a sentence (Islam et al., 2011). By contrast, if the input of the language detector are single words, *f-score* and *accuracy* results are low and not reliable. The output of the LD Toolkit for word input cannot be helpful for the annotator due to the numerous erroneous assignments. Nevertheless, for building a full-formed lexicon, we need to solve the problem of language detection for single unknown words.

In this section, we introduce a model, that processes the output of the LD toolkit and re-assigns a language to each unknown word, reaching higher *f-scores* and *accuracies*. The idea behind this model is that the target word's language is likely to be the same as the language of the sentence, in which this word occurs. In this way, many unknown words in the PL corpus come from editors' comments, containing several sentences in Italian or French. Not only languages assigned to the sentences, in which the word occurs, but also languages assigned to the co-occurring unknown words are important to detect the language of the target word.

The formal model of the calculation looks as follows. Let W be a set of word forms that occur in texts of the corpus C and let $W' \subseteq W$ be the subset of word forms whose language membership is unknown. Further, let S be the multiset of all sentences that occur in texts $x \in C$. Suppose that $f{:}S \rightarrow \mathrm{Pot}(W)$ is a function that maps each sentence $s \in S$ onto the multiset of the word forms that occur in it. The set of all sentences in which any given word form $w \in W$ occurs is defined as

$$S(w) = \{s \in S \mid w \in f(s)\} \tag{3}$$

We proceed by defining the language detection function L_1 that assigns to each word form (known or unknown) its language:

$$L_1{:}\, W \rightarrow \{l_1, \ldots, l_m\} = \mathbb{L} \tag{4}$$

$\mathbb{L}$ is the set of target languages to be detected. Note that we assume m target languages. In our experiments in Section 5, $|\mathbb{L}| = 2$. Note further, that we implement L_1 by means of the LD Toolkit (Section 4.3).

Now we are in a position to make the selection of the target language by means of our lexicon expander model: for any unknown word w it is defined as the language that is

Data: The set of unknown words $W' = \{w_1, \ldots, w_n\}$
Result: The language $\mathcal{L}(w)$ of any word $w_i \in W'$
for $i = 1..n$ **do**
  $\mathbb{L}(w_i) \leftarrow \{l \in \mathbb{L} \mid \exists s \in S(w_i) : l = L_1(s)\}$;
  if $|\mathbb{L}(w_i)| = 1$ **then**
    $\mathcal{L}(w) \leftarrow L_1(w_i)$;
  end
  else
    $\mathcal{L}(w) \leftarrow L_2(w_i)$;
  end
end

Algorithm 1: Lexical Expander language assignment algorithm

most frequently assigned to those unknown words with which w co-occurs in sentences out of $S(w)$. Formally, we define the language detector function $L_2 \colon W \rightarrow \mathbb{L}$ as follows:

$$L_2(w) = \arg\max_{l \in \mathbb{L}} \{|\{w' \in W' \mid \exists s \in S(w) \colon w' \in f(s) \wedge \ L_1(w') = l\}|\} \tag{5}$$

Note that $L_2(w)$ is based on L_1 for guessing the language of unknown words.

In order to enlarge our *tertium comparationis*, we consider a third language detector function $L_3 \colon W \rightarrow \mathbb{L}$. It assigns that language l to an unknown word $w \in W$ that is most frequently assigned to the sentences in which w occurs:

$$L_3(w) = \arg\max_{l \in \mathbb{L}} \{|\{s \in S(w) \mid L_1(s) = l\}|\} \tag{6}$$

Note that L_3 is also based on L_1, but is now applied to whole sentences (as input strings) instead of single words.

Algorithm 1 summarizes the choice between L_1 and L_2 (or L_3).

We applied this algorithm to several bilingual corpora (Section 5). Each time the assignment was done in three steps. The first step was to assign a language to an unknown word by the LD Toolkit. The second step was to assign a language to all the sentences in which the unknown word appears and find the most frequently assigned language. The third step was to assign the language to all the unknown words which co-occur with the target unknown word. Finally, when all the assignment steps successfully passed, we applied our decision algorithm that makes the final assignment.

5 Evaluation

The evaluation was run on four test corpora of various size and topics. (Table 1). Three test sets contain bilingual texts and the fourth test set is multilingual. For bilingual test sets, we used three language pairs: English and Turkish, German and French, OHG and Latin. The first test set is comprised of English Wikipedia articles (e.g. Atatürk, Istanbul etc.), which contained numerous Turkish words. The second one was a collection of German Wikipedia articles, containing French words. The third test set was composed of OHG sentences, containing Latin words. Sentences were manually

Language	Tokens	Sentences	Unknown
German - French	5893	315	460
English - Turkish	14022	724	438
OHG - Latin	1397	217	499
Multiling. German Text	1547	177	344

Table 1: Test corpora

extracted from the OHG corpus[2]. The gold standard includes all the tokens, found in the texts, which are manually annotated according to their language.

For the multilingual test set we used a text by Austrian author Hugo von Hofmannsthal. The basis of this text is an excerpt from his essay *"On the physiology of modern love"* (Hofmannsthal, 1891). It contains long French passages which are mainly quotes of contemporary French authors. Furthermore, Hofmannsthal comments on single sentences and constituents in German. Into this text English translations from the German original have been inserted by the authors in much the same manner, as sentences or constituents. Somewhere in between a ficticious commentary has been added. Additionally, very few sequences, all shorter than three words are Latin. So, except for German, English, French and Latin are also found in the text. In all test sets, all the tokens were manually annotated.

Language	F-Score	Accuracy
German - French	0.40	35.43%
English - Turkish	0.36	38.13%
OHG - Latin	0.79	70.34%
Multiling. German Text	0.37	41.2%

(a) Evaluation of LD toolkit: word level

Language	F-Score	Accuracy
German - French	0.49	49.13%
English - Turkish	0.5	52.51%
OHG - Latin	0.94	96.12%
Multiling. German Text	0.73	74.25%

(b) Evaluation of LD toolkit: sentence level

Language	F-Score	Accuracy
German - French	0.58	53.5%
English - Turkish	0.52	51%
OHG - Latin	0.95	91.78%
Multiling. German Text	0.73	72.96%

(c) Evaluation of the Lexicon Expander

Table 2: Evaluation Results

Table 2a shows the performance of the LD Toolkit for the word-by-word input. The LD toolkit yielded the highest results for the OHG/Latin test corpora and the lowest for the English-Turkish test set among bilingual test sets. The LD toolkit performed poorly on the multilingual test set. Table 2b presents the results of the assignment of the most prominent sentence language to the word. In other words, we calculated the most frequently assigned language of the sentences, where the word occurs, and re-assign this language to the word. *Accuracy* for the OHG text is higher than the one our model reached, but *f-score* is lower in all the cases and for the German text the

[2]TITUS, http://titus.uni-frankfurt.de.

difference in *f-score* between our model and sentence-wise detection is quite big. Table 2c shows results that were reached after the LD Toolkit output was postprocessed by the Lexicon Expander model. For all the test corpora, this model reached higher *f-score* and *accuracy* values than the LD toolkit. The best result is achieved for the OHG text. The average improvement for the bilingual test sets achieved by the postprocessing of the LD toolkit output is around 19% of *accuracy*. As for the multilingual test set, the improvement achieved by the model is even larger. The *f-score* is almost twice as big as the *f-score* of the LD Toolkit and *accuracy* grew for approximately 32%.

6 Discussion and Conclusion

We have shown that the Lexicon Expander as a postprocessing tool improved the performance of language assignment, based on the LD Toolkit, for each of the chosen language pairs. Herein, the Lexicon Expander showed consistent improvement of the *f-score* and *accuracy* values irrespective of the initial performance of the LD Toolkit. Though the language pairs for evaluation were heterogeneous in terms of language relationships (differences of lexical overlap) and text genres, this did not affect the performance of the Lexicon Expander. Further the size of the test corpora and the proportion of unknown words did not influence the improvement. We consider this a hint to the potential of the model in dealing with multilingualism in ancient corpora not only on sentential but also on the lexical level. The improvement of the *f-score* and *accuracy* values suggests that the Lexicon Expander is a first step in developing a functioning toolkit for multi-facetted language detection on various levels (e.g. lexical and sentential levels) and can be of help in saving annotator effort and in preprocessing ancient corpora.

References

Balk, H. (2010). IMPACT annual report 2009, version 1.0. http://www.impact-project.eu.

Bamman, D. and Crane, G. (2009). Structured knowledge for low-resource languages: The Latin and Ancient Greek dependency treebanks. In *Proc. of the Text Mining Services 2009, Leipzig, Germany*. Springer.

Bamman, D., Passarotti, M., Busa, R., and Crane, G. (2008). The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA.

Büchler, M., Heyer, G., and Gründer, S. (2008). eAQUA – bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities – An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Erjavec, T. (2011). 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. Association for Computational Linguistics.

Erjavec, T., Ringlstetter, C., Zorga, M., and Gotscharek, A. (2011). A lexicon for processing archaic language: the case of XIXth century Slovene. In *Proceedings of the International Workshop on Lexical Resources at ESSLLI*.

Gippert, J. (2001). TITUS — Alte und neue Perspektiven eines indogermanistischen Thesaurus. *Studia Iranica, Mesopotamica et Anatolica*, 2:46–76.

Gippert, J. (2006). *Essentials of Language Documentation*, chapter Linguistic documentation and the encoding of textual materials, pages 337–361. Mouton de Gruyter, Berlin.

Gippert, J. (2010a). Manuscript related data in the TITUS project. *ComSt Newsletter*, 1:7–8.

Gippert, J. (2010b). New prospects in the study of old georgian palimpsests. In *Proceedings of the 1st International Symposium on Georgian Manuscripts, October 19–25, 2009, Tbilisi*.

Gleim, R., Mehler, A., Waltinger, U., and Menke, P. (2009a). eHumanities Desktop — an extensible online system for corpus management and analysis. In *5th Corpus Linguistics Conference, University of Liverpool*.

Gleim, R., Waltinger, U., Ernst, A., Mehler, A., Esch, D., and Feith, T. (2009b). The eHumanities Desktop — an online system for corpus management and analysis in support of computing in the humanities. In *Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL 2009, 30 March – 3 April, Athens*.

Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. (2009). Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *AND '09 Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. U., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171.

Hofmannsthal, H. v. (1891). Zur physiologie der modernen liebe. *Die Moderne*.

Hulden, M. (2009). FOMA: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Islam, Z., Mittmann, R., and Mehler, A. (2011). Multilingualism in ancient texts:language detection by example of old high german and old saxon. In *GSCL conference on Multilingual Resources and Multilingual Applications (GSCL 2011), Hamburg, Germany*.

Koster, C. H. A. (2005). Constructing a parser for Latin. In Gelbukh, A. F., editor, *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, volume 3406 of *Lecture Notes in Computer Science*, pages 48–59. Springer.

Mehler, A., Diewald, N., Waltinger, U., Gleim, R., Esch, D., Job, B., Küchelmann, T., Pustylnikov, O., and Blanchard, P. (2011a). Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora. *Leonardo*, 44(3).

Mehler, A., Schwandt, S., Gleim, R., and Jussen, B. (2011b). Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien. *Journal for Language Technology and Computational Linguistics (JLCL)*. Accepted.

Migne, J.-P., editor (1844–1855). *Patrologiae cursus completus: Series latina*, volume 1–221. Chadwyck-Healey, Cambridge.

Passarotti, M. (2000). Development and perspectives of the latin morphological analyser LEMLAT (1). *Linguistica Computazionale*, 3:397–414.

Passarotti, M. (2010). Leaving behind the less-resourced status. the case of latin through the experience of the index thomisticus treebank. In *Proceedings of the 7th SaLTMiL Workshop on the creation and use of basic lexical resources for less-resourced languages, LREC 2010, La Valletta, Malta*, Malta. ELDA.

Smith, D. A., Rydberg-Co, J. A., and Crane, G. R. (2000). The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Sperberg-McQueen, C. and Burnard, L. (2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Waltinger, U. (2010). *On Social Semantics in Information Retrieval*. Phd thesis, Bielefeld University, Germany.

Waltinger, U. and Mehler, A. (2009). The feature difference coefficient: Classification by means of feature distributions. In *Proceedings of the Conference on Text Mining Services (TMS 2009)*, Leipziger Beiträge zur Informatik: Band XIV, pages 159–168. Leipzig University, Leipzig.