



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

1/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

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

Maria Sukhareva

Text Technology Group, Goethe-Universität Frankfurt am Main

5 January 2012



Outline

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

2/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Introduction
- 2 Language Detection
- 3 Lexicon Expander
- 4 Conclusion



Example: Modern Sentences

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

3/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Wenn der Driver beim Link zum Host trappt, muss er mal geupdated werden.
- 2 PHP peut également générer d'autres formats en rapport avec le Web, comme le WML, le SVG, le format PDF, ou encore des images bitmap telles que JPEG, GIF ou PNG.
- 3 French cuisine was codified in the 20th century by Escoffier to become the modern version of haute cuisine; Gastro-tourism and the Guide Michelin helped to acquaint people with the rich bourgeois and peasant cuisine of the French countryside starting in the 20th century.

courtesy: Armin Hoenen



Example: Modern Sentences

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

4/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Wenn der **Driver** beim **Link** zum **Host** trappt, muss er mal geupdated werden.
- 2 PHP peut également générer d'autres **formats** en rapport avec le **Web**, comme le WML, le SVG, le **format** PDF, ou encore des **images bitmap** telles que JPEG, GIF ou PNG.
- 3 French **cuisine** was codified in the 20th century by **Escoffier** to become the modern version of **haute cuisine**; Gastro-tourism and the **Guide Michelin** helped to acquaint people with the rich **bourgeois** and peasant **cuisine** of the French countryside starting in the 20th century.

courtesy: Armin Hoenen



Multilingualism in Modern Texts

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

5/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- The number of multilingual resources available on the web are rising drastically
- Imposing new challenges to NLP researchers
- It is also a challenge for many NLP applications
- There are many language detection toolkits available for modern languages



Example: Ancient Sentences

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

6/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Er uuas miteuuare, also Esaias chat Gaude et letare, Hierusalem, quia rex tuus uenit tibi mansuetus.
- 2 et in anniuersario sancte thiedhilda to then neppenon ande to then almoson ande to themo inganga thero iungereno tue malt
- 3 Tiû grûba uólliu uuazzeres bézeichenet, dáz ér chât Saluum me fac, deus

courtesy: Timothy Price



Example: Ancient Sentences

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

7/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Er uuas miteuuare, also Esaias chat **Gaude et letare, Hierusalem, quia rex tuus uenit tibi mansuetus.**
- 2 **et in anniuersario sancte thiedhilda to then neppenon ande to then almoson ande to themo inganga thero iungereno tue malt**
- 3 Tiû grûba uólliu uuazzeres bézeichenet, dáz ér châ **Saluum me fac, deus**

courtesy: Timothy Price



Multilingualism in Ancient Texts

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

8/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

Sources

- Loan words
- Comments in foreign languages

Patrologia Latina (Mehler et al. 2011)

- 700,000 foreign words (Sukhareva et al. 2011)
- Comments in French, English, German, etc.

Old High German (OHG)

- 9% of words are Latin (Sukhareva et al. 2011)



Multilingualism in Ancient Texts

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

8/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

Sources

- Loan words
- Comments in foreign languages

Patrologia Latina (Mehler et al. 2011)

- 700,000 foreign words (Sukhareva et al. 2011)
- Comments in French, English, German, etc.

Old High German (OHG)

- 9% of words are Latin (Sukhareva et al. 2011)



Multilingualism in Ancient Texts

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

8/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

Sources

- Loan words
- Comments in foreign languages

Patrologia Latina (Mehler et al. 2011)

- 700,000 foreign words (Sukhareva et al. 2011)
- Comments in French, English, German, etc.

Old High German (OHG)

- 9% of words are Latin (Sukhareva et al. 2011)



Language Detection (LD) toolkit (Waltinger and Mehler 2009)

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

9/22

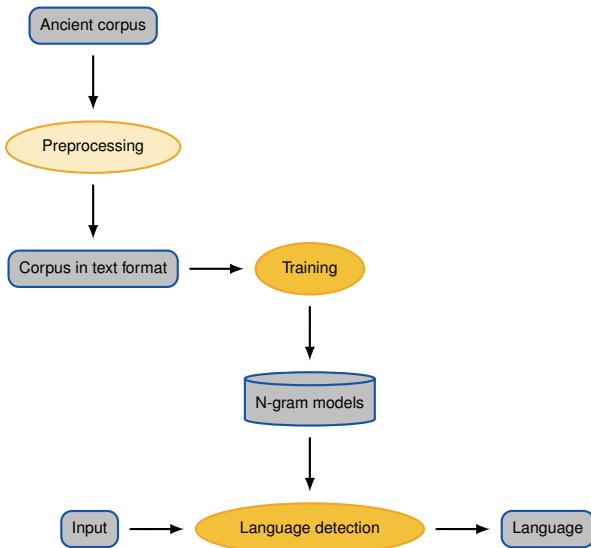
Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion





LD toolkit (Islam et al. 2011)

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

10/22

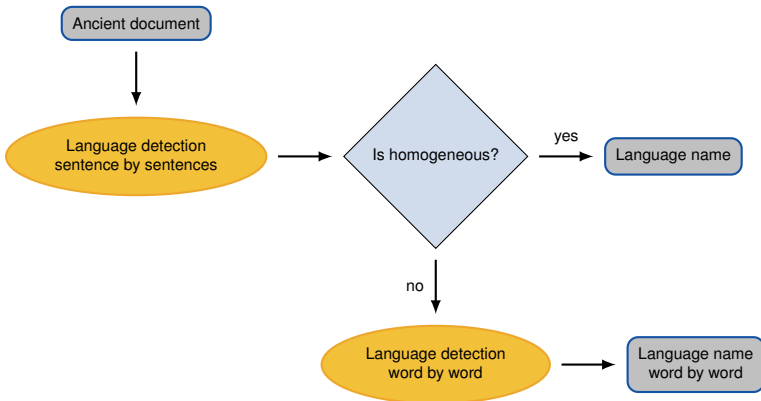
Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion





Evaluation: Test set

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

11/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

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

- English–Turkish test corpus is comprised of English Wikipedia articles (e.g. Atatürk, Istanbul etc.), which contain numerous Turkish words.
- German–French test corpus is comprised of German Wikipedia articles, which contain numerous French words.
- OHG–Latin test corpus is comprised of OHG sentences, which contain Latin words.
- Multilingual test corpus is an excerpt from the German essay of Hugo von Hofmannsthal, containing English, French and Latin insertions.



LD toolkit: Evaluation

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

12/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

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%



Lexicon Expander

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

13/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 An application module for the eHumanities Desktop
- 2 Used to build and annotate lexica
- 3 The LD Toolkit is integrated into it



System Architecture

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

14/22

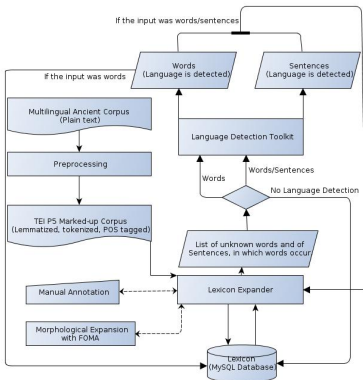
Maria Sukhareva

Introduction

Language Detection

Lexicon Expander

Conclusion



- 1 A multilingual text is preprocessed
- 2 The Lexicon Expander extracts unknown words
- 3 One of three options of language detection is applied
- 4 The results are saved in a MySQL DB
- 5 The user can manually annotate the lexicon or apply morphological expansion



Lexicon Expander

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

15/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

The screenshot shows the Lexicon Reviewer interface. At the top, there are buttons for 'Load Dictionary', 'Download', and 'Create Dictionary'. Below that is a 'Word Index' section with a search filter set to 'Word'. A table displays the following data:

Wordform	Lemma	Language	Pos	AGFL-Lex	Frequency	Autl
sirt	sirt	German	#ADJA		8	sik
iro	iro	German	#ADV		5	sik
unde	unde	Latin	#ADJA		5	sik
dea	dea	German	#ADJO		4	sik
tres	tres	Latin	#ADJA		4	sik
et	et	Latin	#FM		4	sik
sin	sin	German	#VAFIN		3	sik
person?	person?	Latin	#ADJA		3	sik
ierusalem	ierusalem	German	#XY		3	sik
s	s	Latin	#PPER		3	sik
salmun	salmun	German	#ADJA		2	sik
fora	fora	Latin	#VFIN		2	sik
ouh	ouh	Latin	#NN		2	sik
substans?	substans?	Latin	#ADJA		2	sik
spiritus	spiritus	Latin	#ADJA		2	sik
creaturis	creaturis	Latin	#ADJA		2	sik
patrem	patrem	Latin	#FM		2	sik
filium	filium	Latin	#VFIN		2	sik
portio	portio	Latin	#ADJA		2	sik
pilato	pilato	Latin	#ADJA		2	sik
domini	domini	Latin	#ADJA		2	sik
du	du	Latin	#ADJO		2	sik

At the bottom, it says 'Page 1 of 9' and 'Displaying wordforms 1 - 25 of 223'. On the right side, there is a 'Number of sentences' section with an example: '117: in ze ierusalem foru befilhem dku ze sände ist also ouh'.

- The F-score and accuracy are low if the LD Toolkit input is single words
- The Lexicon Expander post-processes the LD Toolkit output, improving the f-score and accuracy
- The Lexicon Expander relies on the language sentences, in which the target word occurs and on the co-occurring unknown words

Figure: The GUI of the Lexicon Expander



Lexicon Expander: The Three Steps

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

16/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- Detect the language of the target word;
- Detect the language of all the sentences in which the target word occurs;
- Detect the language of all the unknown words which co-occur with the target word;



Lexicon Expander

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

17/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

Data: The set of unknown words $W' = \{w_1, \dots, w_n\}$
Result: The language $\mathcal{L}(w)$ of any word $w_i \in W'$

```
for  $i = 1..n$  do
   $L_S(w_i) \leftarrow \{l \in L \mid \exists s \in S(w_i) : l = L_1(s)\};$ 
  if  $|L_S(w_i)| = 1$  then
    |  $\mathcal{L}(w) \leftarrow L_1(w_i);$ 
  end
  else
    |  $\mathcal{L}(w) \leftarrow L_2(w_i);$ 
  end
end
```

Figure: Lexicon Expander Language Assignment Algorithm

where:

- $S(w) = \{s \in S \mid w \in f(s)\}$
- $L: W \rightarrow \{l_1, \dots, l_m\} = L$
- $L_1(w) = \arg \max_{l \in L} \{|\{w' \in W' \mid \exists s \in S(w) : w' \in f(s) \wedge L_1(w') = l\}|\}$
- $L_2(w) = \arg \max_{l \in L} \{|\{s \in S(w) \mid L_1(s) = l\}|\}$



Evaluation: Results

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

18/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

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%

Table: Performance of the Lexicon Expander

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%

Table: Performance of the LD Toolkit



Conclusion

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

19/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- 1 Multilingualism in ancient corpora causes problems for lexicon building
- 2 The Lexicon Expander post-processes the LD Toolkit output and improves f-score and accuracy scores
- 3 This saves annotators efforts and simplifies automatic processing



Reference

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

20/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

- Sukhareva et al. 2011:** Sukhareva, Maria; Islam, Zahurul; Hoenen, Armin; Mehler, Alexander; *A Three-step Model of Language Detection in Multilingual Ancient Texts*, In preparation, 2011;
- Mehler et al. 2011:** Alexander Mehler, Nils Diewald, Ulli Waltinger, Rüdiger Gleim, Dietmar Esch, Barbara Job, Thomas Küchelmann, Olga Pustynnikov, and Philippe Blanchard. *Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora*. Leonardo, 44(3), 2011.
- Islam et al. 2011:** Islam, Md. Zahurul; Mittmann, Roland und Mehler, Alexander ; *Multilingualism in Ancient Texts: Language Detection by Example of Old High German and Old Saxon*, In GSCCL conference on Multilingual Resources and Multilingual Applications (GSCCL 2011), 28-30 September, Hamburg, Germany, 2011.
- Waltinger and Mehler 2009:** Ulli Waltinger and Alexander Mehler, *The feature difference coefficient: Classification by means of feature distributions*, In Proceedings of the Conference on Text Mining Services , Leipziger Beiträge zur Informatik: Band XIV, pages 159–168. Leipzig University, Leipzig.



Acknowledgement

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

21/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

This work is funded by LOEWE Digital- Humanities project in the Goethe-Universität, Frankfurt.



digital humanities

Goethe-Universität Frankfurt | Technische Universität Darmstadt | Freies Deutsches Hochstift



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

22/22

Maria Sukhareva

Introduction

Language
Detection

Lexicon Expander

Conclusion

Thank you!