ANALYSIS OF PART-OF-SPEECH TAGGING OF HISTORICAL GERMAN TEXTS

DATECH 2017

Markus Paluch, Gabriela Rotari, David Steding, Maximilian Weß, Maria Moritz, Marco Büchler

June 1 2017, *University of Göttingen*

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

INTRODUCTION

RESEARCH QUESTION

RQ: Should POS-Taggers be trained on a certain epoch/period?

POS-TAGGING

POS-Tagging: The process of marking up the words in a text to a particular part of speech (tag).

POS-TAGGING EXAMPLE

POS-Tagging: The process of marking up the words in a text to a particular part of speech (tag).

Word / Token	Tag	Wordclass
Money	NN	noun
does	DOZ	does
not	*	negation
smell	VB	verb
.	.	punctuation

Not all words correspond to a single wordclass.

mobile	JJ	adverb
mobile	NN	noun

WHAT IS A POS-TAGGER?

POS-Tagger trained on dataset X: A computer model which learned to perform POS-Tagging on texts in X.

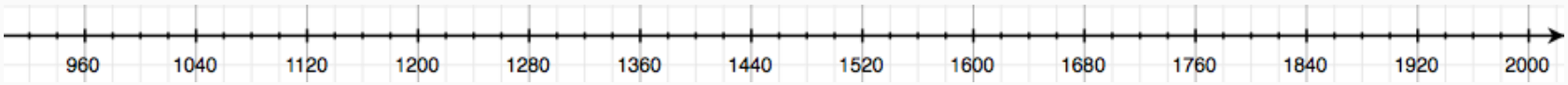
POS-TAGGING ON TEXTS OF DIFFERENT LANGUAGES

German ≠ English, it is known that:

A German trained tagger processing English texts **performs badly** and vice versa.

Historical German ≠ Contemporary German, we ask:

Does a tagger trained on contemporary German processing historical German texts **performs badly** and vice versa?



RESEARCH QUESTION

RQ: Should POS-Taggers be trained on a certain epoch/period?

DATA

German Text Archive (Deutsches Text Archiv, DTA)[1]

- comprises 1598 texts
- dating from 1050 to 1926

1. Berlin-Brandenburgische Akademie der Wissenschaften. Deutsches Textarchiv. http://www.deutschestextarchiv.de/. Online; accessed 24-May-2016.

DATA PRESENTATION

Period		Texts	Tokens
Baroque	1600-1720	76	9,935,705
Romanticism	1810-1840	264	15,470,398
Modernism	1880-1920	87	6,027,221

Table 1: Datasets for the experiment

EXPERIMENTAL SETUP

USED TAGGERS

Included POS-Tagger algorithms[1]:

- **Unigram** 
- Hidden Markov Model (**HMM**) 
- Conditional Random Field (**CRF**) 
- **Perceptron** 

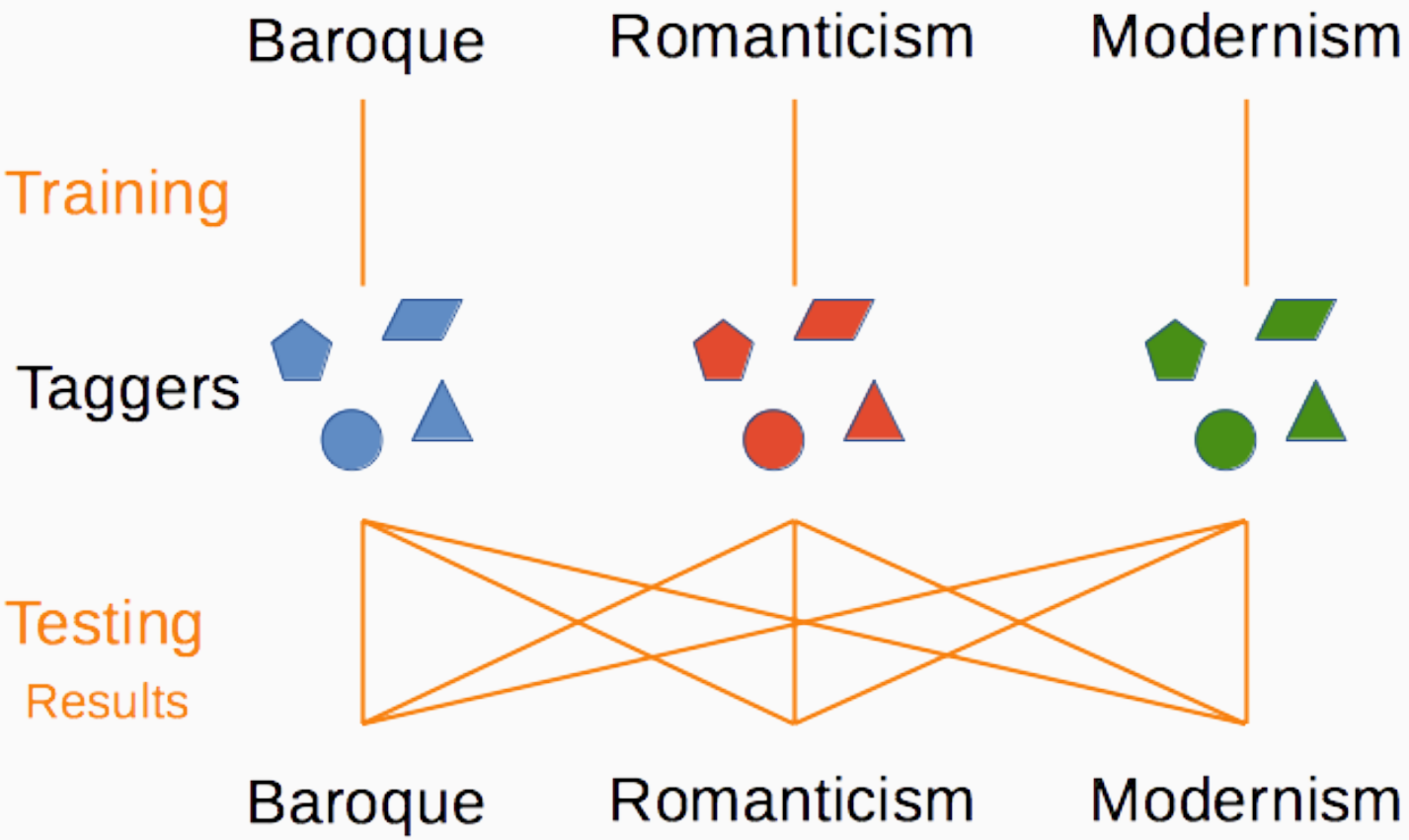
1. All used algorithm implementations are from the natural language toolkit (NLTK)

PROCEDURE

Procedure:

1. Training of taggers on data
2. Testing of taggers (Results)

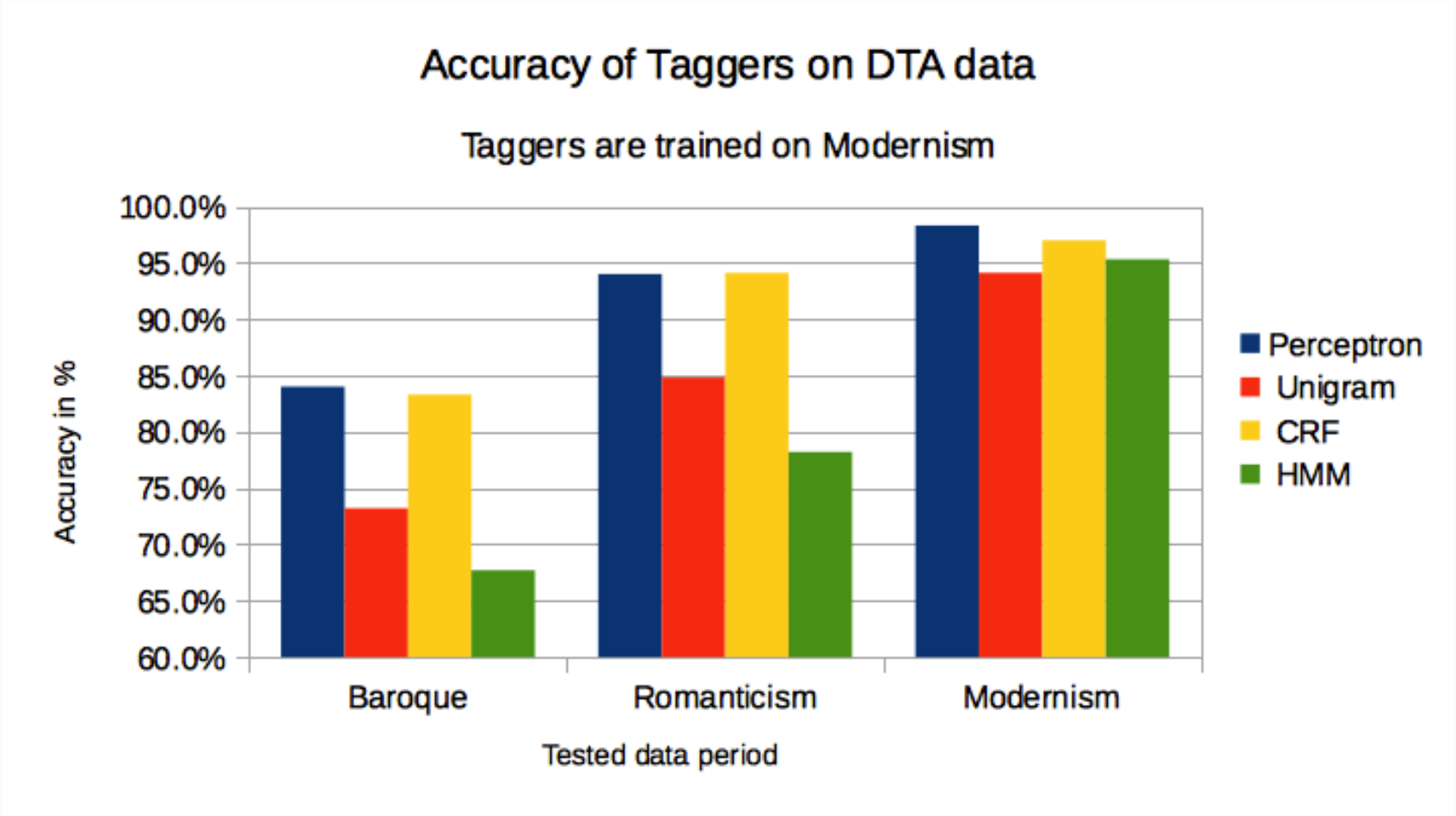
METHODOLOGY



RESULTS

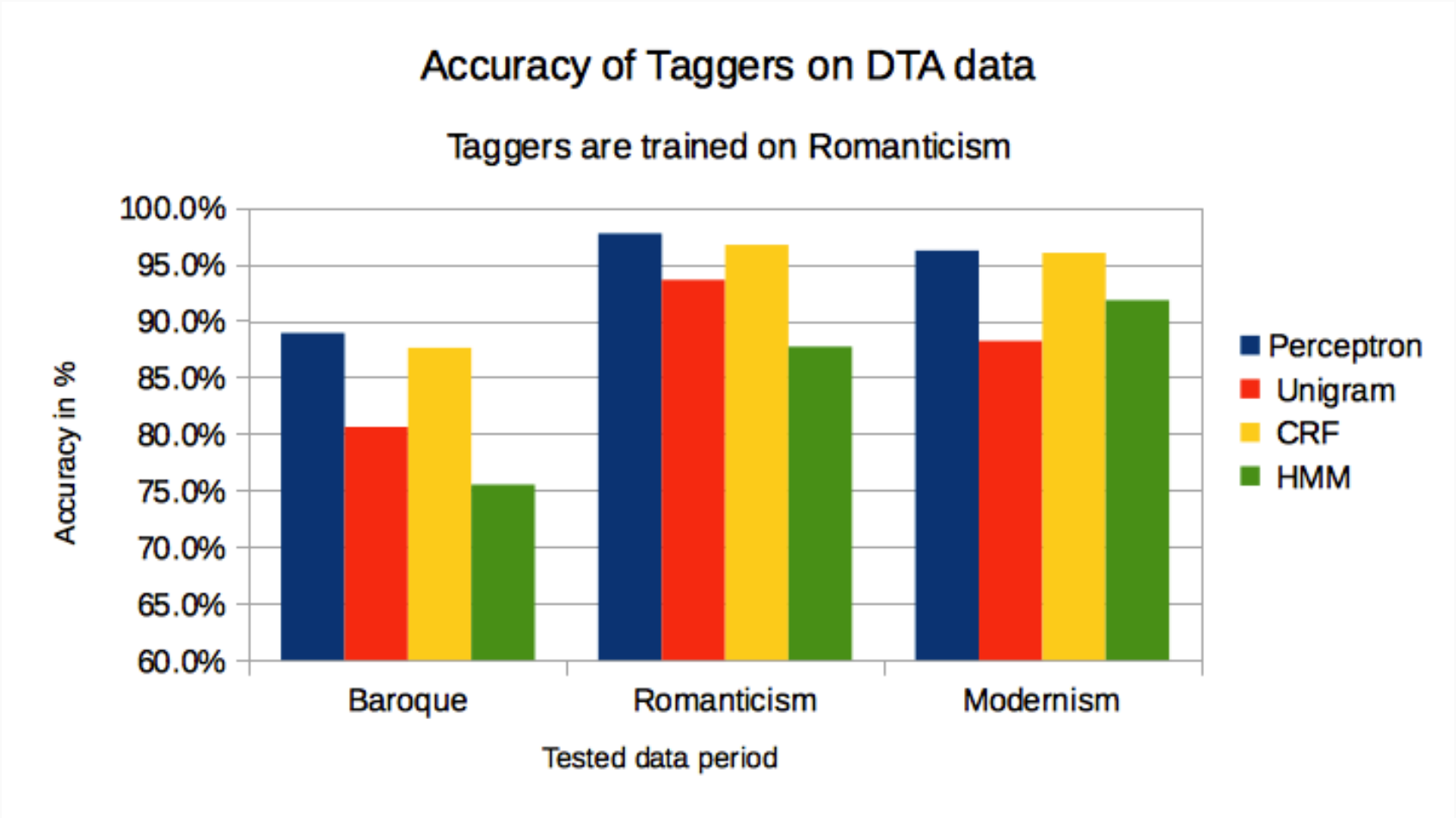
RESULTS OF MODERNISM TAGGERS

Modernism



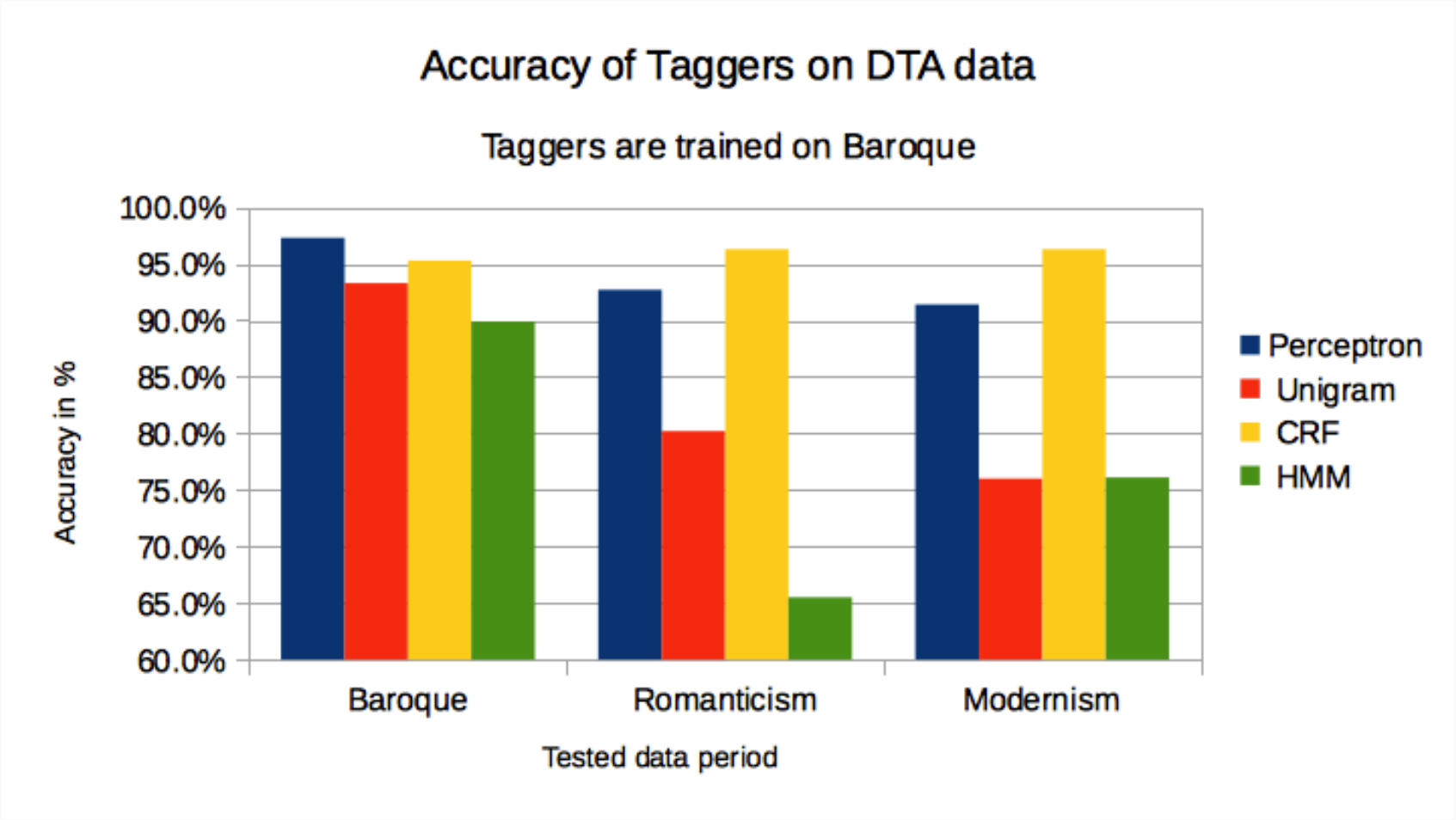
RESULTS OF ROMANTICISM TAGGERS

Romanticism



RESULTS OF BAROQUE TAGGERS

Baroque



COMPARING DTA AND HANDTAGGING RESULTS

We handtagged about 300 tokens of 1 text per period.

What happens if a tagger trained on non goldstandard data (DTA) is tested against goldstandard data (handtagging)?

Taggers	Accuracy	
trained and tested on	DTA	Handtagging
Modernisim	94.1%-98.3%	91.7%-95.6%
Romanticism	87.7%-97.7%	93.6%-96.8%
Baroque	89.9%-97.3%	88.1%-90.5%

CONCLUSION

CONCLUSION

1. Using a POS-Tagger trained on a different period of the same language can dramatically decrease its performance!
    - Higher time differences between periods increase the performance decrease.
2. DTA POS-Tags for Baroque are more erroneous than POS-Tags of Romanticism or Modernism on our handtagged examples.

RQ: Should POS-Taggers be trained on a certain epoch/period? Yes!

CONTACT

Visit us


http://www.etrap.eu


contact@etrap.eu

Stealing from one is plagiarism, stealing from many is research
(Wilson Mitzner, 1876-1933)

SPONSORED BY THE

LICENCE