Multi Tier Annotation Search (MTAS) & Nederlab

INCEpTION workshop Technische Universität Darmstadt

March 2018

Matthijs Brouwer

Meertens Institute

Royal Netherlands Academy of Arts and Sciences

KNAW Humanities Cluster

KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Meertens Instituut

Multi Tier Annotation Search - MTAS

Search and analysis on annotated and structured text

Textual data with structure and annotations : POS, NE, sentences, paragraphs, …

- Based on Apache Lucene
- Can be used as a Solr plugin
- Scalable, support distributed search
- Combined with metadata

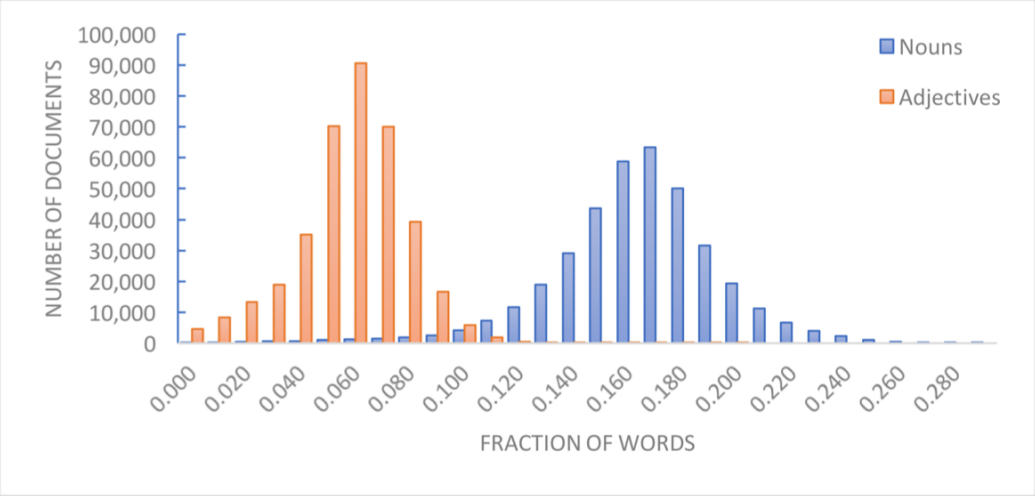
Features MTAS

- Support multiple formats annotated text
- Extensive configuration options
- Combined search/analysis on metadata and annotated text
- Annotations on non-single positioned items
- Store hierarchical relations
- CQL query language (Corpus Query Language)
- Scalable, support distributed search

Features MTAS

Statistics

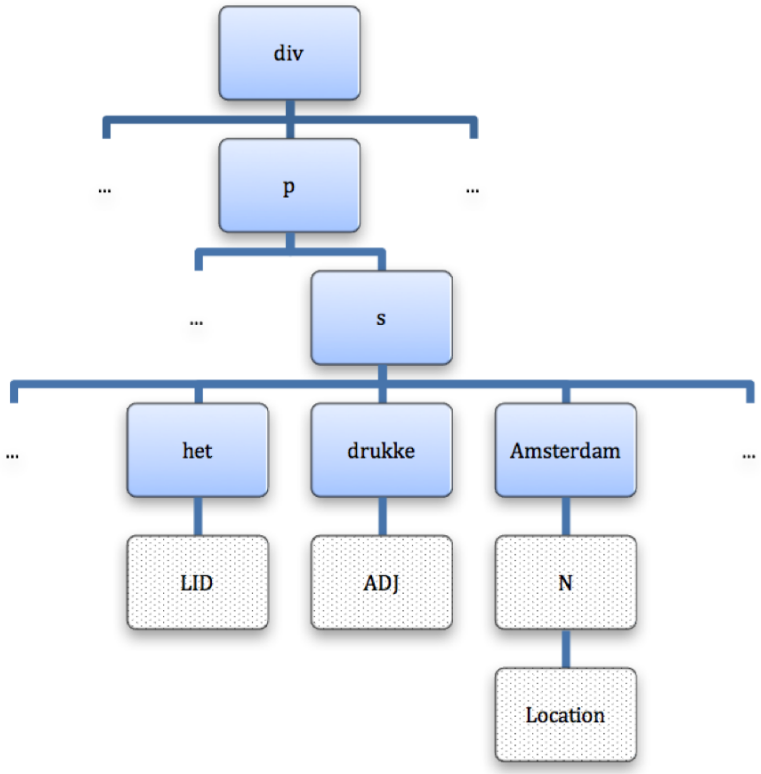
[pos = “N”]



Type	Value
sumsq	9435447605455
populationvariance	23686821.654429
max	618900
sum	451966125
kurtosis	1723.7663754426
standarddeviation	4866.917377324
n	375367
quadraticmean	5013.6407199361
min	0
median	209
variance	23686884.757698
mean	1204.0646220859
geometricmean	0
sumoflogs	-Infinity
skewness	25.092673770659

Features MTAS

Keyword In Context representations



[pos="LID"] [pos="ADJ"] "Amsterdam"

Find documents / positions

Use Forward index Mtas

Features MTAS

Termvector or **frequency lists** over all documents (matching a condition)

most frequent terms containing 5 letters and ending with e

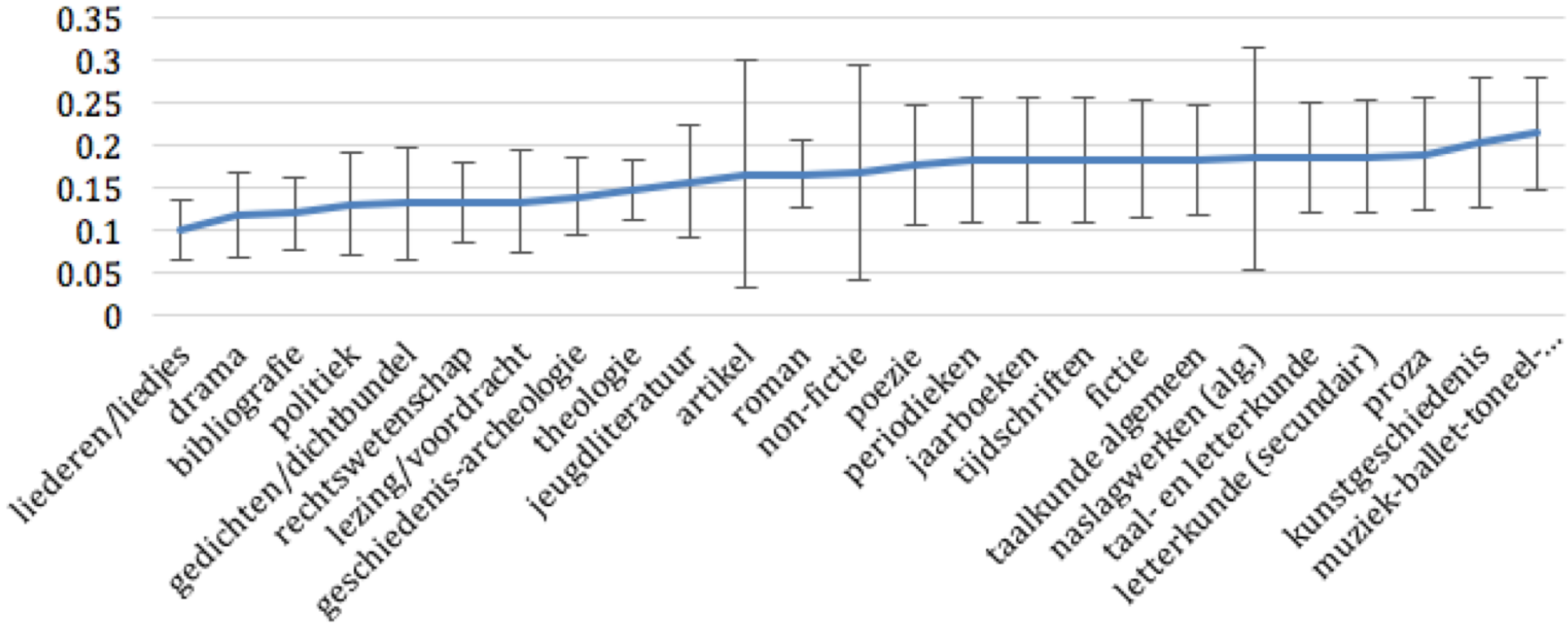
	Documents	Total	Mean	Median	Max	Mean	Median	Std. deviation	Kurtosis	Skewness
Term	Frequency					Relative frequency				
welke	4,170,151	12,242,628	2.94	1	3,996	0.0032	0.0021	0.0046	223.4	10.9
zijne	2,628,115	7,541,324	2.87	1	4,523	0.0030	0.0020	0.0036	56.1	5.29
hunne	2,291,385	5,237,717	2.29	1	3,994	0.0025	0.0016	0.0031	63.8	5.53
goede	2,268,787	4,081,615	1.80	1	1,318	0.0027	0.0015	0.0043	709.2	14.2
einde	1,954,052	3,351,655	1.72	1	893	0.0021	0.0012	0.0031	4932	27.3

Features MTAS

$$\frac{\#\ [\ pos = \text{“ADJ”}\ ]\ [\ pos = \text{“N”}\ ]}{\#\ [\ pos = \text{“N”}\ ]}$$

Faceting

Facet : genre



Features MTAS

Group results

<entity="loc" /> within (<p/> containing ("simon" "stevin"))

Location	Documents	Hits
brugge	67	144
leiden	61	117
gent	26	95
antwerpen	48	89
amsterdam	33	85
brussel	19	50

Nederlab

- One of the primary use cases for MTAS
- Aims to bring together all digitized texts relevant to the national heritage of the Netherlands, the history of Dutch language and culture

	Total
Documents	15.7 million
Words	9.6 billion
Annotations	36.5 billion

Annotations

- Documents for Nederlab are (automatically) tokenized and annotated before inclusion
- Occasionally updates on tokenisation and annotation

- We would be able to support **user annotations** and store this on word position level
- However not included in search results and analysis
- And connection to original text cannot be guaranteed on updates

Future work

- User upload documents
  - Search & analysis on these documents
  - Profit from combination with existing corpora

- Syntactic tree search, Penn Treebank format
- Ranking and sorting of results
- Topic modelling

THANK YOU

Nederlab

www.nederlab.nl

MTAS

github.com/meertensinstituut/mtas/

meertensinstituut.github.io/mtas/

E-mail: matthijs.brouwer@meertens.knaw.nl