# Mining Paper Catalogues

A Multilingual Solution to Reduce Verbose Fields to Consistent Terminology

Felix Kußmaul

University of Cologne

Dr Tim Evans

UNIVERSITY *of* York

31 August 2017

EAA Maastricht 2017

# MOTIVATION

# Data Source

**Form 23**  Conical cup with smooth vertical rim
Konische Schale mit glattem Steilrand
Coppa troncoconica con orlo verticale
Coupe tronconique à rebord vertical lisse

Conical cup representing the further evolution of Form 22. The floor is now always flat or biconical (meeting the wall at a sharp angle on the inside), usually with a low foot.

23.1: Plain tapering rim, inclined slightly inwards, sometimes bearing applied decoration.
23.2: Rim with flat outer face bearing applied decoration bounded above and below by simple convex mouldings; inner face plain or with a groove at lip.

### Production
Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

### Date
Subform 23.2 belongs to the second and third quarters of the first century A.D.: it is common in the South Stoa deposits at Corinth and at Pompeii. Subform 23.1 is less readily datable as it may occur as a simplified version of Form 22 or Form 23: other features of the vessel (e.g. foot-profile, decoration) may provide a clearer indication of date than the shape of the rim.

### Distribution
Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

### References
23.1.1 Karthago K 78/172a, unpublished. Stamp L.MA, O.-C. –. Italy.
23.1.2 Berenice B210.2. Anepigraphic stamp. Italy.
23.2.1 Corinth 1973 pl.84,70. Stamp CAMVRI, O.-C. 397. Arezzo.
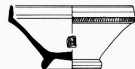23.2.2 Berenice B216.2. Italy.

### Other findspots
23.1 Asciburgium, Bologna, Bolsena, Conimbriga, Köln, Luni, Magdalensberg, Ordona, Pollentia, Roma.
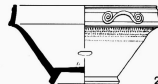23.2 Not separately listed.

### Concordance
23.1: Goudineau 20c; 25a; 37a. – Berenice B210.2.
23.2: Goudineau 40. – Barocelli 11. – Berenice B216. – Hayes 23.
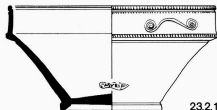Pieces described as Haltern 9 sometimes belong to this form.

P.M.K.



**Figure 1:** Sample from *Conspectus* catalogue.

# Oh dear!

### Problem

Running texts contain a lot of *irrelevant information* (for machine processing).

This makes database lookups <span style="color:orange">extremely inefficient</span>.

## What we **have**:

### Production

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

### Distribution

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

UNSTRUCTURED DATA

## What we **want**:

```
{
  "form": "23.1",
  "origin": "Italy",
  "decoration": "none",
  "occurs": "uncommon"
},
{
  "form": "23.2",
  "origin": "Italy, not Padana",
  "occurs": "Mediterranean region;
            North-Italy"
}
```

STRUCTURED DATA

## What we have:

**Production**

Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

**Distribution**

Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

UNSTRUCTURED DATA

## What we want:

```
{
    "form": "23.1",
    "origin": "Italy",
    "decoration": "none",
    "occurs": "uncommon"
},
{
    "form": "23.2",
    "origin": "Italy, not Padana",
    "occurs": "Mediterranean region;
               North-Italy"
}
```

STRUCTURED DATA

## What we **have**:

### Production
Subform 23.1 is probably made in many parts of Italy; examples in Padana ware do not show applied decoration. Subform 23.2 is made in Italy but apparently not in the Padana region.

### Distribution
Subform 23.2 is very common throughout the Mediterranean region, with sporadic examples found in the North and in North-Italy; Subform 23.1 is relatively uncommon.

**UNSTRUCTURED DATA**

## What we **want**:

```
{
    "form": "23.1",
    "origin": "Italy",
    "decoration": "none",
    "occurs": "uncommon"
},
{
    "form": "23.2",
    "origin": "Italy, not Padana",
    "occurs": "Mediterranean region;
               North-Italy"
}
```
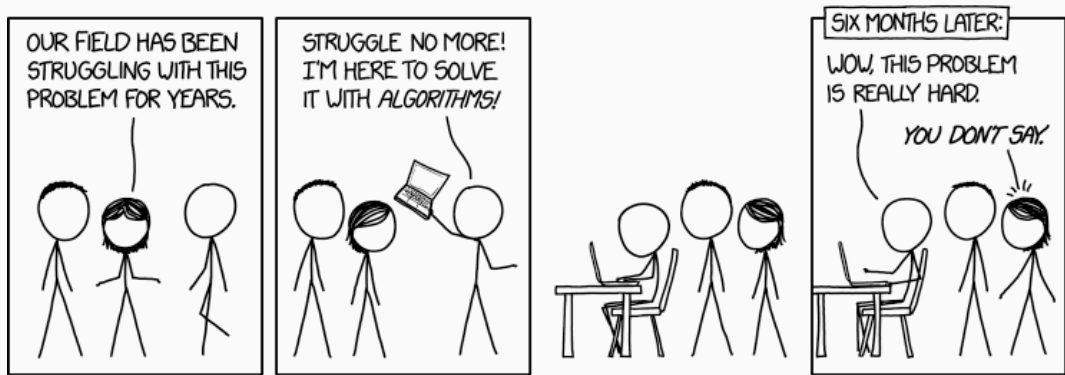
**STRUCTURED DATA**

Figure 2: I can relate to this. [Source: xkcd.com/1831]

# TEXT MINING: THEORY

# Classification

## Definition: Text Mining

**Text Mining** is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- Information extraction
- ...

# Classification

## Definition: Text Mining

**Text Mining** is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- Information extraction
- ...

# Classification

## Definition: Text Mining

**Text Mining** is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- Information extraction
- …

# Classification

## Definition: Text Mining

**Text Mining** is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- ...

# Classification

## Definition: Text Mining

**Text Mining** is a general term covering several different ideas, e. g.:

- Information retrieval
- Statistical analysis
- **Information extraction**
- ...

# Information Extraction

### Definition: Information Extraction (IE)

*"[IE] is the task of automatically extracting structured information from unstructured […] documents."*

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is really super difficult and often inaccurate.

**Definition: Information Extraction (IE)**

*"[IE] is the task of automatically extracting structured information from unstructured [...] documents."*

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is really super difficult and often inaccurate.

### Definition: Information Extraction (IE)

*"[IE] is the task of automatically extracting structured information from unstructured [...] documents."*

Some other facts about Information Extraction:

- Computer scientists have a hard time with IE (for over 30 years now!)
- IE is really super difficult and often inaccurate.

## DISCLAIMER

We neither *can* nor *do* provide a perfect solution or perfect results.

Furthermore, this project is still work in progress.

# Five Steps

1. Tokenisation and Sentence splitting
2. Lemmatisation
3. Part-of-speech-tagging (POS)
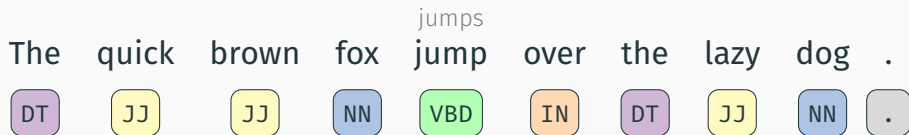4. Named entity recognition (NER)
5. Relation Extraction

# POS-Tagging



Figure 3: POS-tagging examples after lemmatisation.

The quick brown fox jump over the lazy dog .

jumps

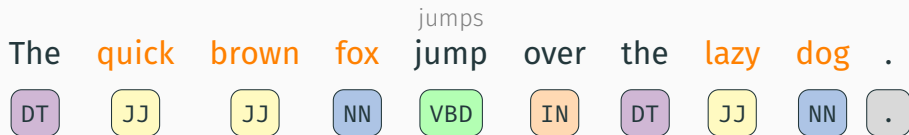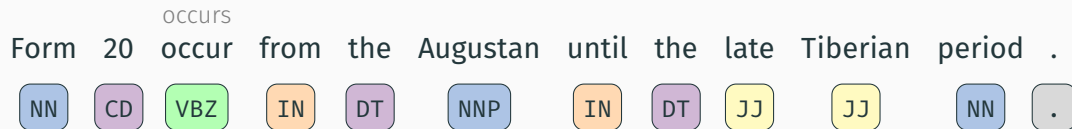| DT | JJ | JJ | NN | VBD | IN | DT | JJ | NN | . |

Figure 3: POS-tagging examples after lemmatisation.

Figure 4: POS-tagging examples after lemmatisation.

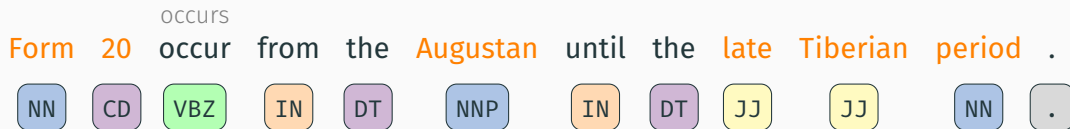Figure 4: POS-tagging examples after lemmatisation.

# Relation Extraction

| Subject | Relation | Object |
|:---:|:---:|:---:|
| quick brown fox | jump over | lazy dog |
| Form 20 | occur | Augustan |
| Form 20 | occur | late Tiberian period |

# TEXT MINING: PRACTICAL

# Tools

presenting different tools here

## Adapting the NER

Stanford CoreNLP only recognises 8 entities types:

| | |
|---|---|
| PERSON | DATE |
| ORGANIZATION | TIME |
| LOCATION | MONEY |
| PERCENT | MISC |

So we have to add the custom type FORM. Adjusting DATE also necessary.

nuthin yet

# MULTILINGUALISM

# Mining Paper Catalogues

A Multilingual Solution to Reduce Verbose Fields to Consistent Terminology

Felix Kußmaul

University of Cologne

Dr Tim Evans

UNIVERSITY *of* York

31 August 2017

EAA Maastricht 2017