University of Antwerp
Faculty of Arts

Modelling the past

The use of digital text analysis techniques for historical research

Outline

- 2 case studies:
  - (Fe)male voices on stage: finding patterns in lottery rhymes of the late medieval and early modern Low Countries, with and without AI
    - Collaboration with Marly Terwisscha van Scheltinga and Jeroen Puttevils
    - Cultural history
  - Named Entity Recognition and Classification for Early Modern English
    - Collaboration with Patrick Quick (internship + MA thesis)
    - NLP
- "Digital History": on machines and manuscripts

(Fe)male voices on stage

Finding patterns in lottery rhymes of the late medieval and early modern Low Countries, with and without AI

Marly Terwisscha van Scheltinga, Sara Budts and Jeroen Puttevils

Broader context and overview

- Project aim
  - Explore the self-identification of women in the Early Modern Low Countries through lottery rhymes
  - PhD research of Marly Terwisscha van Scheltinga

- This article
  - Classify Early Modern Dutch lottery rhymes based on the gender of their author
  - To appear in Low Countries Historical Review (BMGN) 2024 (1)

Early modern lotteries

- Delft, 1564: Cornelis Janssen comes to buy 6 lottery tickets
  - In the clerk's registry:
    - "Per Delft Cornelis Janssen scipper vanden boechige anden turffmart"
    - 6 tickets
    - "Cornelis Janssen scipper van delft mit sijn zes kinderen hadde hij tgroote lot ten zoude hem niet hinderen"

      "Cornelis Janssen, skipper of Delft with his 6 children, if he won the jackpot, it would't impede him"
  - In the lottery rhyme container:

Cornelis Janssen scipper van delft mit sijn zes kinderen hadde hij tgroote lot ten zoude hem niet hinderen

X 6

  - In the prize container:

NIET / PRIZE

X 6


1592

Women's self-identification in the early modern period

- Women's voices in early modern sources
  - Many administrative sources (tax rolls, testaments, court records, …)
    - Women relatively absent, defined in relation to men
  - Lottery rhymes: self-identification
- Relative freedom of women in early modern Low Countries

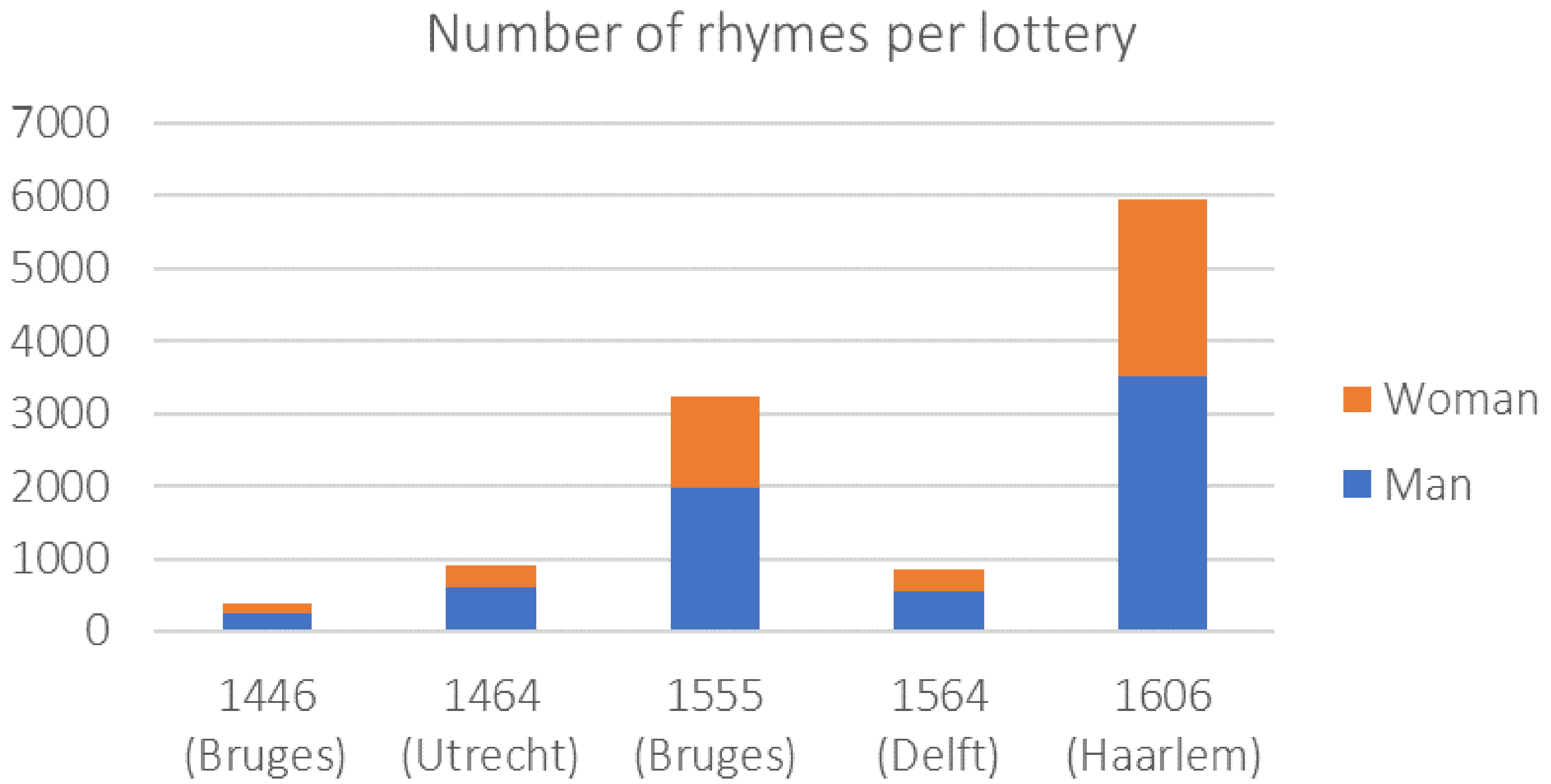
Research questions

- Discourse analysis
  - Discursive patterns rather than morphosyntax
  - Relation to patriarchal norms (cf. Howell 2019)
    - Reshaping them?
    - Conforming to them?
    - Challenging them from within?

Which patterns can we see in the lottery rhymes of women from the middle of the fifteenth to the beginning of the seventeenth century and to what extent were these similar to or different from those of their male counterparts?

Dataset

- 11 332 lottery rhymes



Gender marking in the lottery rhymes

- With self-identification (with overt gender markers)
  - name, personal pronouns, profession or other gendered references

Lenaert Adriaenssen, builder at the nieuwe langendijck, if he won the jackpot he would be rich.
Maijken, spinster in the Three Cups; Mary, virgin pure, wants to grant her the jackpot.

- Without self-identification (without overt gender markers)

A E I O U the five vowels, what will they get me?
If God grants me luck, I will share it with the poor.
I put in to win.

Approach (1)

- **GysBERT**
  - Manjavacas & Fonteyn (Universiteit Leiden)
  - Trained on DBNL and Delpher (7.1B tokens; 1500-1950)
    - DBNL = small, but clean
    - Delpher = large, but noisy
  - Why GysBERT?
    - Used to historical language material (spelling and morphology)
    - Can find discursive patterns beyond the level of the single word
      - Templates, lottery rhymes
      - We're looking for discursive differences

Approach (2)

- **Train classifiers:**
  - Non-gender related: geographical, diachronic, social
  - Gender-related

- **Technical implementation:**
  - Hugginface's Transformers library ("BertForSequenceClassification")
  - Weights inverse to class frequency
  - Hyperparameter tuning through 5 WandB sweeps:
    - Batch size, no. epochs and learning rate that optimised Macro avg. F1
  - Trained on Colab's GPUs
  - Model with highest macro avg. F1 picked for prediction on validation data

Approach (3)

- **Manual reconstruction of classification cues:**
  - Procedure:
    - Sort validation data by descending classification probability for M/F
    - Search topsamples for recurrent features (e.g. presence of name of certain rhyme)
    - Annotate entire validation set for said feature
    - Test if feature is significantly more present for one gender than the other
  - Vs. automated measures (SHAP, LIME)
    - Too computationally expensive
    - Assumes tokens of a given sentence to be (locally) linearly separable
    - We're interested in discursive features:
      - Multi-words
      - Including non-consecutive chunks

Non-gender classifiers (architecture and results)

Classification	No. classes	Dataset size	Batch size	No. epochs	Learning rate	Mac. avg. F1	Random baseline (macr. avg. F1)
geographical	9	2 510	16	5	3.97E-05	0.569	0.078
diachronic	5	11 338	32	3	4.57E-05	0.843	0.161
social	2	5 850	32	6	2.43E-05	0.552	0.435

Geographical variation

Classification	No. classes	Dataset size	Batch size	No. epochs	Learning rate	Mac. avg. F1	Random baseline (macr. avg. F1)
geographical	9	2 510	16	5	3.97E-05	0.569	0.078

- Geographical reach of a lottery could be large

  e.g. Bruges 1555 had participants from The Hague, Utrecht and Amsterdam

- Relatively small influence of geography:

  e.g. rhymes from Utrecht and Holland put in at the 1555 lottery in Bruges had more in common with the other (Flemish and Brabantian) rhymes of that lottery than with the Holland rhymes of the 1564 Delft lottery

Diachronic variation

Classification	No. classes	Dataset size	Batch size	No. epochs	Learning rate	Mac. avg. F1	Random baseline (macr. avg. F1)
diachronic	5	11 338	32	3	4.57E-05	0.843	0.161

- Clear development in time:
  - Bruges 1446 and Utrecht 1464: many identification-only rhymes
  - Haarlem 1606: only 38% had any identifiers at all
- Variation in rhyming templates, e.g.:
  - Bruges 1555: 'Jesus van Nazarenen' ('Jesus of Nazareth) ~ 'verlenen' ('to grant')
  - Haarlem 1606: '

University of Antwerp
Faculty of Arts

Social variation

Classification	No. classes	Dataset size	Batch size	No. epochs	Learning rate	Mac. avg. F1	Random baseline (macr. avg. F1)
social	2	5 850	32	6	2.43E-05	0.552	0.435

- Bulk buyers (> average) vs. small quantity buyers (< average)
- Difficult to classify
- Bulk buyers were less likely to mention their names:
  - Preferred to show off knowledge / send moralising message?
  - Didn’t want their name to be read out loud so often?

Gender variation – the easy way

Jhesus davids zone int ghemeene Gheeft Katelijnken Gheleijns tgrootste lot tot haren deele
Cathelijne Ghuus zelden zaghende es naer den upperprijs vraghende
Lijsbet Luchten wijf uuijt den haghe Hadde lijever tgrote lot bij nacht als bij dage
Lijsken Fobeleijne inde drapstrate per Mechelen
Jhesus van nazareenen wil Tanneken de Smit tgrootste lot verleenen
Maeijcken Fevers Nam gaerne tgroot lot ghegeven
Lijsbeth inde munt
Maijken van Smaelden Sal thoocxste lot halen
Betken de kousmaijcstere tot delft inde pepersteech Zij woudt dat zij tgrote lot mit Jesus creech
Barbele Jan Zuevels wijf jn de Eechoutstrate

Jan Lemens in Sinte Peeters goidshuijs Hij hadde geren den hoochsten prijs thuijs
Simon Ameberghen te middelburch ghebooren Had hij het hoocxste lot Hij en waer niet mede verlooren
Jan Backele Hadde hij het hoochste lot Hij waer wel tevree
Jan Willems de hantschoemaker bij sint jacops kercke Hadde hij het hoochste lot Hij en soude niet veel wercken
Claes Wellen Crech hij thoochste lot Hij sout wel tellen
Jannijn Bultijeu inde hoocxstrate int paradijs Hadde hij het hoocxste lot Hij waer wel wijs
Adriaen Haghens inde corte nieustraete inde drije mollen Hadden hij het hoocxste lot Het soude hem wel bollen
Pieter Maertijnssen tapper int oest eijnden inden aeckeren boem alias scram Hij hadde liever tgroette lot dan een vetten ram
Pieter Stiers tantwerpen aende wilde zee inden gulden visscher Hadde hij thoochste lot Hij soude den trecker prijsen
Thomas Vermeeren Had hij het hoochste lot Hij soudt wel begheere

Masking the overt gender markers

- (1) By type of gender marker; (2) uniformly

  - Lenaert Adriaenssen, builder at the nieuwe langendijck, if he won the jackpot he would be rich.

<NAME>, <OCCUPATION> at the nieuwe langendijck, if <PRON> won the jackpot <PRON> would be rich
<IDENTIFIER> at the nieuwe langendijck, if **<IDENTIFIER>** won the jackpot **<IDENETIFIER>** would be rich

  - Maijken, spinster in the Three Cups; Mary, virgin pure, wants to grant her the jackpot.

<NAME>, <NOUN> in the Three Cups; Mary, virgin pure, wants to grant <PRON> the jackpot.
<IDENTIFIER> in the Three Cups; Mary, virgin pure, wants to grant **<IDENTIFIER>** the jackpot.

Gender classifiers (architecture and results)

Classification	Dataset size	Batch size	No. epochs	Learning rate	Macr. avg. F1	Random baseline (macr. avg. F1)
Gender in rhyme (no mask)	5 330	16	5	4.61E-05	0.967	0.497
Gender in rhyme (masked)	5 265	16	6	1.88E-05	0.589	0.497
Gender in rhyme (masked ID)	5 255	8	2	3.39E-05	0.577	0.497
Gender not in rhyme	4 600	16	5	1.43E-05	0.526	0.490
All data of 1555	3 235	16	4	4.01E-05	0.619	0.494
All data of 1606	6 040	8	4	3.89E-05	0.542	0.496

Gender related variation

- **Identification markers**
  - Men were more likely to mention their occupation
    - Higher range of occupations (and higher in status)
  - Women were more likely to mention their marital status

  BUT rare in both cases and inverse trend as time progresses
  - Women were more likely to mention their name

- **Themes and tropes**
  - Women appealed more to divine entities
  - No difference in mentions of charity (despite womens' reputation of caregivers)

Gender related variation

- **Template preferences**
  - Women: 'Jong van jaren' ('young of years') ~ 'bewaren' ('preserve')
  - "I sold X and brought the money into the lottery":
    - Used about equally by women and men, but gendered variation in X:
      - Women sold textiles and food; men sold tools, animals and gaming objects (e.g. marbles and knucklebones)
  - Did women use more templates altogether?
      - Women use more templates than men *proportionally*
      - Men were leading the shift away from rhymes with (only) identification markers

  -> Women were more conservative in writing lottery rhymes

Conclusions (historical)

- **Structured variation in the lottery rhymes?**
  - Above all diachronic: rhymes evolved as a genre
  - Low frequency of occupation and marital status <-> administrative sources
  - In terms of gender: the difference are subtle, but significant:
    - Women appealed more often to divine entities
    - Women didn’t hesitate to mention their names in public
      (even when this was no longer the trend)
    - Women adhered more to templates
      (but gave their own spin to them)
  - Female rhymes might have been more conservative, but they were a ‘license to speak’ regardless

Conclusions (methodological)

Could we have found all patterns manually? Yes, probably

Would we have found all patterns manually? Absolutely not

- **Division of labour**
  - Computers are good at finding patterns but struggle to interpret/contextualize them
  - People are good at interpreting patterns but struggle to keep track of them

  -> let NLP assist us in highlighting which parts of the dataset require more human attention and contextualisation
- **Manual annotation/reconstruction of classification cues = bottleneck**
  - Thoughts/advice more than welcome!

NER for Early Modern English

NERing the Johnson Letters (1542-1552)

Patrick Quick and Sara Budts

Aims (1)

- **Broader context: Back2TheFuture**
  - Research project on future thinking among European merchants (1400-1800)
  - Johnson correspondence is part of the corpus
- **NER needed for**
  - Corpus exploration: who is mentioned?
  - Reconsruction of their physical world: where did they go?
  - ! Reconstruction of their timescape
    - How far in time did they look ahead ?
    - Individual variation ?
    - Different temporal outlooks for different aspects of their lives?
    - How did they structure their time? (Day/month/year; holidays; markets)

Aims (2)

- **Named Entities to extract**
  - Names
  - Locations
  - Dates and holidays (absolute markers of time)
    - e.g. 4 december 1551; Saint Bartholmew day; 2nd of this present
  - Relative temporal references
    - E.g. yesterday, in 4 days, often, every once in a while, ...
  - Markets and fairs
    - 4 main Brabantian fairs (sinxon, paessche, bames and cold) + local fairs
  - Divine entities
    - God, Jesus and all saints; “Lord” and “father”
  - Nations
    - E.g. Hollanders, Ynglyshemen
  - Price
    - E.g. 70li 11s 8d Fl.

Named Entity Recognition

= extracting named entities from running text

- E.g. names of persons, places, organisations, ...

- **Many different approaches:**
  - Rule-based search (e.g. regular expressions)
  - Non-neural supervised learning:
    - Hidden Markov Models (just word and transition probabilities)
    - Conditional Random Fields (flexibility thanks to features)
  - Deep learning, supervised:
    - Convolutional Neural Networks (+ CRFs = neuro-CRF)
    - Transformers
  - Unsupervised learning:
    - Clustering, but of limited use

NER for historical texts

- **4 complicating factors (Ehrmann et al. 2023)**
  - Document type and domain variety
    - Gaps between domains exist for present day data, but no info on historical texts
  - Input noise
    - OCR, HTR (manuscripts), OLR (layout), a manual transcriber’s mistakes
  - Language dynamics
    - Variation in spelling, morphology, meaning, ...
    - Named entities are specific to (historical) context
  - Resource Availability
    - Relative lack of training data

The Johnson Letters

- **John Johnson and his network (1542-1552)**
  - John, Otwell and Richard Johnson
  - Wool and fell trade between Northampton, London, Calais and Antwerp
- **The correspondence**
  - 881 handwritten letters
  - Some outgoing, some incoming
  - 77 different letter writers
  - Haphazardly preserved (patchy coverage)
  - Handwritten, but transcribed in 1953 and recently OCRd and corrected
  - Mainly Early Modern English, some Early Modern French

Examples (after transcriptions, before OCR)

JOHN JOHNSON TO ANTHONY WHITE

Jhesus anno 1546, the tirde in January, at Ticford.

Mr. White,

I comende me unto you, and praie I maie be the same to Mistris Fayrey youre mother. Accordinge unto your request, I did sende unto Mr. Kyrkham youre lettre, and I wrot hym(1) I had disbursyd so moche to your mother as by your letter ye willid him to paie me; but he sent me awurde that he had certain billis of his, and wold not paie his monney without he myght receave them, saing further that the morowe after Twelfte daie he wolde be at London and satisfie you, and therfore ye maye at his thither comynge provide to geyt youre monney. Nevertheles, I wolde it shuld not seme unto Master Kirkham but that I had dysbursyd the monney to your mother, bycause ye wrote me so, and also bycause I declarid the same in my lettre unto him, so that yf he shuld perseave the contrary, it myght be occasion to make him conseave displeasure towardes me, and that I wold not have. Wherfore I praie you when ye speke with him, declaire that I was dysapointid of my monney by meanes he paide me not, and also yf he paie you shortely, cause my brother Otwell to receave it, bycause it maie seme to him the rather to be trew. Thus in hast I comyt you to youre Lorde.

By youres,

John Johnson.

WILLIAM HOWHAM TO JOHN JOHNSON

Jhesus.

Syr,

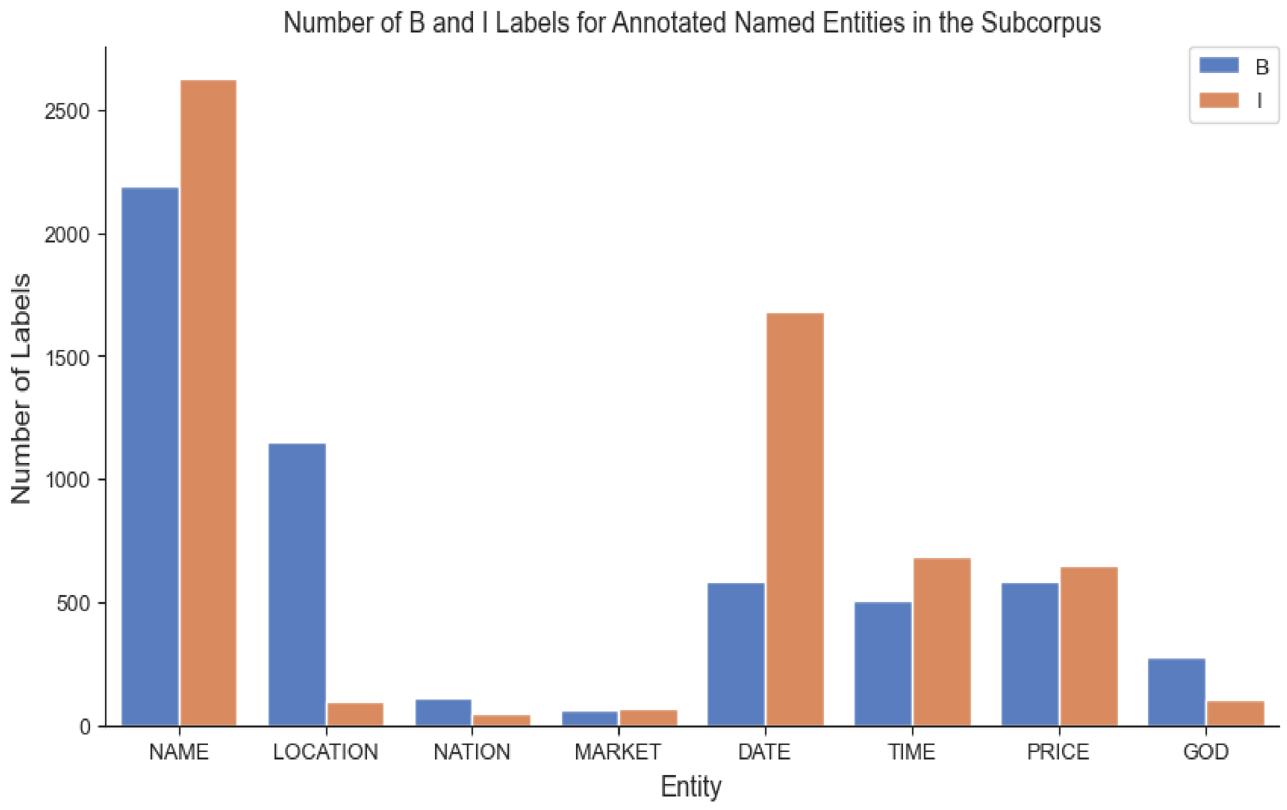
In my best maner I recomend me unto you, and to my maystrysse, hertely desseyryg off God your good welfare. Syr, I undyrstand that you intend be the grac off God to be ressedent and dewllyg in this contre. Syr, I have a dowtur wyche hathe bene at servys iij or iiij yere in the contre, and brokyne with all werkes [for] a womane to do, and now off laytt sche ys comyne wh[ome after] hyre terme. Yf you be unpurvyd, I wold be glad yt [might be] you to have hyre; and be my feythe, yff I dyd kn[ow any] vysse or condeschns be hyre, sche schuld nat [come]; and I pray you off answere hereoff, for unto soche [time as I hear] frome you, sche schall nat be fest with no mane. I wold t[hat it would] plesse you to have hyre. No more, but Owre Lord send you off [His grace], amen. At Peturborow on Fast Tewsseday,(1) anno xiij.

Be your that I cane,

Wyllyam Howham.

Training data

- **BIO tagging (Jurafky & Martin 2023)**
  - No nested labels
  - 159 letters were manually tagged



Related work

- **Low-resource NER**
  - Data augmentation (e.g. replacement with identically labelled words) (Dai & Adel 2020)
  - Distant supervison to label unlabelled data (external data sources)
  - **Fine-tuning** (depends on gap between domains) (e.g. Torge et al. 2023)
- **NER with historical data (Ehrmann)**
  - Adapting the data by **removing the noise:**
    - **Spelling normalisation (Baron & Rayson 2008; Bolmann 2019)**
    - **Create many potential normalisations for the model to choose from (Hämäläinen et al. 2018)**
  - **Adapting the system** to the noise

Data preprocessing

- **Tokenisation**
  - Splitting on whitespace
  - Punctuation removed
- **Spelling normalisation**
  - By means of VARD
  - Combination of parameter settings
  - 30 versions of the corpus
  - 21 of which were unique

Original Location Entities

B label	Count
London	211
Callais	101
Calleis	65
Glapthorne	52
Cales	26

Normalised Location Entities

B label	Count
London	214
Calais	115
Underpay	77
Collies	69
Glapthorne	52

Original Nation Entities

B label	Count
the	32
Frenche	7
Englisshe	6
Hollanders	5
The	5

Normalised Nation Entities

B label	Count
the	32
English	12
French	9
Oleanders	7
Frenchmen	5

Experiments

- **Baselines**
  - Lexical lookup (name, location, market, god, nation, time) + regex (date, money)
  - SpaCy's default EntityRecognizer for English
- **Non-neural**
  - CRF
- **Neural**
  - Bert-base-NER (present-day English model for NER)
  - hmBERT (historical multilingual model for NER, 1800-1900)
  - MacBERTh (generic model for historical English, 1450-1950)

Conditional random field

- **Features**
  - token, POS, some character-level substrings, lowercase token, capitalised?, BOS?, EOS?
  - Preceding and following word: token, POS, casing
- **4 stages:**
  - Preliminary: constrained CRF model ran on every subcorpus
  - Best subcorpus: full training + test of CRF
  - Undersampling: lines without positive labels removed from training data
  - Combined sampling: 2 best corpora (if different F1) + undersampling
    - ~ data augmentation

Deep learning

- **Finetuning 3 different base models**
- **Only for name, location, date, time and price**
- **All named entities modelled individually**
- **2 different corpora:**
  - Original corpus (no spelling normalisation)
  - Best performing subcorpus for CRF
- **80-10-10 training-validation-test split**

Name

	Original	Normalised
Baselines		
Spacy	0.50	0.54
Lexical lookup	0.54	0.56
Conditional Random Fields		
Best subcorpus	0.9455	
Undersampling	0.9448	
Combined sampling	0.9486	
Neural Models		
Bert-base-NER	0.9527	0.9435
hmBERT	0.9519	0.9327
MacBERTh	0.9265	0.9197

Location

	Original	Normalised
Baselines		
Spacy	0.50	0.60
Lexical lookup	0.43	0.46
Conditional Random Fields		
Best subcorpus	0.7147	
Undersampling	0.7104	
Combined sampling	0.6931	
Neural Models		
Bert-base-NER	0.8433	0.7234
hmBERT	0.7960	0.7290
MacBERTh	0.7522	0.7649

Date

	Original	Normalised
Baselines		
Spacy	0.26	0.29
Lexical lookup	0.34	0.34
Conditional Random Fields		
Best subcorpus	0.9020	
Undersampling	0.8840	
Combined sampling	0.8782	
Neural Models		
Bert-base-NER	0.9095	0.9029
hmBERT	0.9325	0.9170
MacBERTh	0.8865	0.8755

Time

	Original	Normalised
Baselines		
Spacy		
Lexical lookup	0.33	0.34
Conditional Random Fields		
Best subcorpus	0.6581	
Undersampling	0.6903	
Combined sampling	0.6557	
Neural Models		
Bert-base-NER	0.6747	0.7337
hmBERT	0.7373	0.7281
MacBERTh	0.6944	0.7226

Price

	Original	Normalised
Baselines		
Spacy	0.01	0.02
Lexical lookup	0.68	0.63
Conditional Random Fields		
Best subcorpus	0.8759	
Undersampling	0.8768	
Combined sampling	0.8759	
Neural Models		
Bert-base-NER	0.9290	0.9067
hmBERT	0.9414	0.9201
MacBERTh	0.8989	0.9099

God

	Original	Normalised
Baselines		
Spacy		
Lexical lookup	0.37	0.36
Conditional Random Fields		
Best subcorpus	0.9205	
Undersampling	0.9363	
Combined sampling	0.9368	
Neural Models		
Bert-base-NER		
hmBERT		
MacBERTh		

Nation

	Original	Normalised
Baselines		
Spacy	0.04	0.19
Lexical lookup	0.44	0.46
Conditional Random Fields		
Best subcorpus	0.7489	
Undersampling	0.8189	
Combined sampling	0.8391	
Neural Models		
Bert-base-NER		
hmBERT		
MacBERTh		

Market

	Original	Normalised
Baselines		
Spacy		
Lexical lookup	0.49	0.45
Conditional Random Fields		
Best subcorpus	0.7828	
Undersampling	0.7609	
Combined sampling	0.8261	
Neural Models		
Bert-base-NER		
hmBERT		
MacBERTh		

Conclusion

Entity	Best model	F1 macro
Name	bert-base-NER$^{\text{original}}$	0.9527
Location	bert-base-NER$^{\text{original}}$	0.8433
Nation*	Combined sampling	0.8391
Market*	Combined sampling	0.8261
Date	hmBERT$^{\text{original}}$	0.9325
Time	hmBERT$^{\text{original}}$	0.7373
Price	hmBERT$^{\text{original}}$	0.9414
God*	Combined sampling	0.9368

- **Neural models outperform CRF models**
  (but small difference for name)
- **Normalisation does not yield better results**
- **Experimenting with sampling techniques pays off**
- **Larger neural models that are specialised for NER are better than generic models for historical text:**
  - Training data size > in-domain training data

Wrap-up

On historians and machines

On machines and manuscripts

- Differences between branches of historical research
  - Economical history vs. cultural history
  - E.g. Cliometrics (1960s, revival in 1990s

- “Linguistic turn” (1970s) + today’s NLP
  - Potentially very fruitful combination:
    - Importance of discourse
    - New ways of studying discourse at scale
  - Hasn’t really taken off yet (<-> historical linguistics)
  - Is gaining momentum:
    - e.g. BMGNs special issue on digital history
    - e.g. Jo Guldi's "The Dangerous Art of Text Mining" (2023)

On machines and manuscripts

- **Frequent issues**
  - Spelling variation
  - Relatively small datasets
  - Tailored pre-trained models are rare

- **Potential solutions**
  - Normalise data to match present-day language (?)
  - Leave data as they are and domain adapt present-day language model (!)
  - Leave data as they are and fully train custom language model (?)

On machines and manuscripts

- Issues that remain
  - Interpretability is key!
    - Why did the models produce the output they produced?
      - Expressiveness vs. Explainabiliity
      - On discourse level!
    - Especially for historians!
      - "historical method"
      - Corpus balance
      - Critical stance towards the sources
    - Cf. "The Dangerous Art of Text Mining" (Guldi 2023)

Thank you

For your attention