Improving part-of-speech annotation for socially diverse historical corpora

Lidia Pivovarova, Janine Siewert, Lassi Saario-Ramsay, Samuli Kaislaniemi, Akseli Kettunen, Inga Kokkonen & **Tanja Säily**

Introduction

- POS-tagging Late Modern English (LModE):
  average accuracies of 94–96% using e.g. VARD + CLAWS
  - Baron (2011a,b); Schneider et al. (2016:258);
    Saario et al. (2021:124)
- However, **accuracy varies by gender and education level**
  → problematic for historical sociolinguistics
  - Saario et al. (2021:119–120)
- Bigger corpora → manual correction unfeasible
  → better methods needed

Promising approach: Language models

- Kulick et al. (2022): accuracies up to 98.30% using pre-trained ELMo embeddings
  - *Early English Books Online* (published texts), simplified Penn tagset
- Manjavacas & Fonteyn (2022:5-6): historically pre-trained MacBERTh performs better than other BERT models (c. 90%)
  - *Penn-Helsinki Parsed Corpus of Early Modern English*, Penn tagset
- No work on social variation in tagging accuracy, CLAWS

Our pilot study

Part of the project “New Methods for Developing Diverse Corpora” (DEDICO, University of Helsinki, 2025-)

- Use the CLAWS7 tagset
- LModE personal letters, socially representative sample
- Comparison of approaches:
  - Normalization (VARD+manual) > statistical **CLAWS tagger**
  - **Shallow neural models**: Stanza, MaChAmp
  - **Large language models**: Qwen2.5 family (+ GPT-5.5 in Codex)
- Best method? Social variation in tagging accuracy?
  - Quantitative + qualitative evaluation

Training data

- *British National Corpus Sampler* (BNC)
  - 2M words
  - 1990s
  - Written and spoken
  - CLAWS7, hand-corrected
- *Corpus of Early English Correspondence Extension* (CEECE)
  - Personal letters from the long 18th century, based on published original-spelling editions, CLAWS7 tagged (TCEECE)
  - Compiled for historical sociolinguistics
    → social metadata, representativeness
  - Gold standard sample: 15 hand-corrected letters, ~5K words

Test data: CEECE sample 2

Letter ID	Sender Name	Gender	Rank	Year	Word Count
BENTHAJ_016	Jeremy Bentham	Male	Professional	1772	316
BLOMEFI_034	Francis Blomefield	Male	Clergy (Lower)	1748	349
BURNEY_040	Charles Burney	Male	Professional	1784	321
CARTER_015	Elizabeth Carter	Female	Clergy (Lower)	1739	362
DUKES_089	Charles Lennox	Male	Nobility	1746	361
HADDOC2_008	Richard Haddock	Male	Professional	1709	310
HATTON2_048	Elizabeth Hatton née Scroggs	Female	Gentry (Lower)	1690	393
PAUPER_011	Ann Clark	Female	Other	1768	215
PINNEY_056	Nathaniel Pinney	Male	Merchant	1706	353
PRIDEA2_035	Humphrey Prideaux	Male	Clergy (Upper)	1710	323
SANCHO_034	Ignatius Sancho	Male	Other	1779	388
SWIFT_033	Mary Butler nèe Somerset	Female	Nobility	1723	374
SWIFT_036	John Carteret	Male	Nobility	1724	311
TWINING_006	Elizabeth Twining née Smythies	Female	Clergy (Lower)	1764	325
YOUNG_026	Edward Young	Male	Clergy (Lower)	1730	332

Examples from ‘other non-gentry’

- Ann Clark (female, 1768):
  - “he after words indeverd to woork at abrookous But got ahurt in his leg which was histeth and left me and my famaley in the gratest destress”
- Ignatius Sancho (male, 1779):
  - “You have missed the truth by a mile - aye and more - It was not neglect - I am too proud for that - it was not forgetfulness, Sir, I am not so ungrateful - [...]”

Ignatius Sancho

Analysis: methods and results

Statistical baseline: normalization + CLAWS

- First normalized by VARD (Baron 2011a,b), then by hand
  - Not completely; targeted most frequent variants, abbreviations, etc. (Saario & Säily 2020:§3)
  - E.g. I receved your Cind Letter → I received your Kind Letter
- Tokenized and POS tagged by the CLAWS tagger
  - I_PPIS1 received_VVD your_APPGE Kind_NN1 Letter_NN1
- Tagging accuracy on the test dataset 94.28%

Shallow neural models

- Comparison between Stanza (Qi et al. 2020) and MaChAmp (van der Goot et al. 2021)
- Three training setups:
  - Training only on BNC
  - Pre-training on BNC, fine-tuning on CEECE
  - Training only on CEECE
- Best overall model with 91.97% accuracy: MaChAmp pre-trained on the BNC and fine-tuned on the CEECE
- Relatively low accuracy of Stanza (81.89% at most) due to issues with annotating punctuation

Large language models

- Instruction-following models from Qwen2.5 family
- Fine-tuned using LORA method
- Fine-tuning on a bigger BNC sample is essential for model to learn CLAWS7 tagset and specific instructions
  - Difficulty with ditto tags (in_II31 terms_II32 of_II33)
- Best-performing model is QWEN2.5-**72B**-Instruct but difference with **7B** is only 0.7pp

eval	Fine-tuning dataset			
	-	CEECE only	BNC only	BNC → CEECE
strict	72.58	84.34	92.15	**92.26**
–ditto	72.90	84.53	**93.38**	93.35

Comparison of accuracy by document and overall

Document	CLAWS	Shallow	LLM
BENTHAJ_016	**95.28**	94.44	93.89
BLOMEFI_034	**96.46**	94.95	92.93
BURNEY_040	**94.53**	92.45	88.80
CARTER_015	**96.89**	95.34	96.37
DUKES_089	**95.97**	93.45	94.21
HADDOC2_008	**94.13**	91.90	93.58
HATTON2_048	**93.41**	86.82	91.14
PAUPER_011	77.38	76.02	**84.16**
PINNEY_056	**88.37**	86.43	**88.37**
PRIDEA2_035	**96.43**	91.76	95.05
SANCHO_034	**94.94**	93.32	85.22
SWIFT_033	95.16	93.70	**97.58**
SWIFT_036	**96.47**	92.94	94.41
TWINING_006	**95.99**	92.25	95.45
YOUNG_026	**95.43**	93.91	91.37
Overall	**94.28**	91.76	92.26

- Comparison of the best shallow model and the best LLM against the baseline (CLAWS)
- CLAWS is still the best overall (thanks to normalization)
- LLM outperforms CLAWS on three letters
- LLM differs from CLAWS the most on lower-class letters:
  - PAUPER_011 +6.79pp
  - SANCHO_034 -9.72pp

Comparison of accuracy by contextual variables

Document	CLAWS	Shallow	LLM
Sender Gender			
- Female	93.13	89.97	**93.73**
- Male	**94.83**	92.62	91.55
Sender Rank			
- Nobility	**95.83**	93.39	95.48
- Gentry	**93.41**	86.82	91.14
- Clergy	**96.24**	93.68	94.20
- Professional	**94.65**	92.92	92.01
- Merchant	**88.37**	86.43	**88.37**
- Other	**89.51**	87.97	84.90
Year of writing			
- Pre-1740	**94.54**	91.56	93.49
- Post-1740	**93.98**	92.00	90.82

- CLAWS prevails here as well
  - Social variation in tagging accuracy: worse accuracy for women + lower classes
- However, LLM does slightly improve the accuracy of women's letters (+0.6pp)

Comparison of accuracy by major tag groups

	CLAWS		Shallow		LLM	
Tag group	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Punctuation	99.83	**99.66**	98.81	99.49	**100.00**	83.08
A-	**100.00**	**98.93**	96.58	96.38	99.14	98.29
C-	**95.28**	**93.80**	90.11	87.08	93.19	91.99
D-	94.65	**95.68**	89.20	84.86	**96.67**	94.05
I-	**96.82**	**96.31**	95.42	95.25	93.09	92.27
J-	93.96	88.30	88.04	86.17	**94.89**	**92.20**
M-	92.65	86.30	**93.24**	**94.52**	92.96	90.41
N-	89.81	89.63	89.70	89.43	**95.48**	**91.54**
P-	99.66	**99.14**	98.25	96.40	**100.00**	97.95
R-	**90.65**	**90.35**	75.29	82.32	82.46	86.17
V-	92.18	93.97	88.54	90.95	**98.38**	**94.55**
Misc	97.00	**89.81**	95.96	87.96	**97.47**	89.35

- LLM does well in precision, CLAWS in recall
- What model you should use depends on the word class you are studying
- Note on calculation principles:
  - E.g. RR instead of RG counts as an error here
  - The figures would be higher if we only looked at the first letter of the tag

Examples from Ann Clark (female, other, 1768)

- “he **after words** indeverd to woork at abrookous”
  - Correct: after_RT21 words_RT22 (afterwards_RT)
  - All models tag separately (e.g. after_II words_NN2)
- “send me **sum** relefe **so That** with industry imay bee able to git my bread”
  - “sum” (DD). CLAWS: NN1, Shallow and LLM: DD
  - “so that” (CS21 CS22). CLAWS: CS21 CS22, Shallow: RR DD1, LLM: CS CS
- “All my frends thare is **Ded** which mad me not Go”
  - Correct: JJ
  - CLAWS: NP1, Shallow: VM, LLM: JJ

Example from Ignatius Sancho (male, other, 1779)

“You have missed the truth by a mile - aye and more - It was not neglect - I am too proud for that - it was not forgetfulness, Sir, I am not so ungrateful - it was not idleness, the excuse of fools - nor hurry of business, the refuge of knaves - It is time to say what it was.”

- Extensive use of dashes/hyphens (-) as punctuation
- Punctuation marks should be tagged like any token, but the LLM did not tag them (although it had a similar letter from Sancho already in the training data)

Conclusions

Summary of results

- Best method: CLAWS, but only because of normalization
  - Social variation in tagging accuracy:
    worse accuracy for women + lower classes
- LLM promising: improves tagging of women's letters, even the pauper letter
  - Also better for some word classes, precision
  - This was a 72B-parameter open-source LLM
    → also try commercial LLMs

Sneak preview: CLAWS vs. GPT-5.5 in Codex

Document	CLAWS	GPT-5.5
BENTHAJ_016	95.28%	**96.67%**
BLOMEFI_034	**96.46%**	95.20%
BURNEY_040	94.53%	**97.92%**
CARTER_015	**96.89%**	96.11%
DUKES_089	95.97%	**97.73%**
HADDOC2_008	94.13%	**96.93%**
HATTON2_048	93.41%	**95.91%**
PAUPER_011	77.38%	**91.40%**
PINNEY_056	88.37%	**93.07%**
PRIDEA2_035	96.43%	**97.53%**
SANCHO_034	94.94%	**96.56%**
SWIFT_033	95.16%	**98.31%**
SWIFT_036	96.47%	**97.94%**
TWINING_006	95.99%	**96.26%**
YOUNG_026	95.43%	**97.21%**

- Overall accuracy:
  CLAWS 94.28% vs. **GPT-5.5 96.32%**
  - Even the pauper letter >91%
    (but still the least accurate)
- No fine-tuning, just prompting
  - We thank Jukka Suomela for conducting the experiment
- GPT-5.5 even found some inconsistencies in our gold standard…

Conclusion

- Commercial LLM seems to provide the best performance
- Why bother with non-commercial models?
  - We have control over the process, e.g. fine-tuning for improvement
  - More stable, long-term solution
  - Free (access to a supercomputer helps; CSC - IT Center for Science)
  - Smaller models use fewer resources
  - Data not shared with a company
- Future work
  - Try out AI-based normalization as a prior step
  - Benchmark different LLMs

References

Baron, Alistair. 2011a. VARD 2. Computer program. http://ucrel.lancs.ac.uk/vard/

Baron, Alistair. 2011b. *Dealing with spelling variation in Early Modern English texts*. Lancaster University dissertation. https://eprints.lancs.ac.uk/id/eprint/84887/

BNC = The BNC Consortium. 2005. *The BNC Sampler, XML version*. Oxford University Computing Services. http://www.natcorp.ox.ac.uk

CLAWS. Computer program. Developed by UCREL at Lancaster University. http://ucrel.lancs.ac.uk/claws/

Kulick, Seth, Neville Ryant & Beatrice Santorini. 2022. Parsing Early Modern English for linguistic search. Allyson Ettinger, Tim Hunter & Brandon Prickett (eds.), *Proceedings of the Society for Computation in Linguistics 2022*, 143–157. ACL. https://aclanthology.org/2022.scil-1.12/

Manjavacas, Enrique & Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, NLP4DH. https://doi.org/10.46298/jdmdh.9152

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Saario, Lassi & Tanja Säily. 2020. *POS tagging the CEECE. A manual to accompany the Tagged Corpus of Early English Correspondence Extension (TCEECE)*. Helsinki: VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/tceece_doc.html

Saario, Lassi, Tanja Säily, Samuli Kaislaniemi & Terttu Nevalainen. 2021. The burden of legacy: Producing the *Tagged Corpus of Early English Correspondence Extension* (TCEECE). *Research in Corpus Linguistics* 9(1): 104–131. https://doi.org/10.32714/ricl.09.01.07

Schneider, Gerold, Marianne Hundt & Rahel Oppliger. 2016. Part-of-speech in historical corpora: Tagger evaluation and ensemble systems on ARCHER. Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 256–264. Ruhr-Universität Bochum. https://konvens.org/proceedings/2016/

TCEECE = *Tagged Corpus of Early English Correspondence Extension*. Annotated by Lassi Saario & Tanja Säily. Spelling standardized by Mikko Hakala, Minna Palander-Collin, Minna Nevala, Emanuela Costea, Anne Kingma & Anna-Lina Wallraff. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily & Anni Sairio at the Department of Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/

van der Goot, Rob, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf & Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 176–197. ACL. https://doi.org/10.18653/v1/2021.eacl-demos.22

Large language models: prompt

You are a CLAWS7 part-of-speech tagger for 18th-century letters.

Your job:

- Read the input text and output token-level CLAWS7 tags.
- Keep original token order.
- Output one tag table and nothing else.

- Hard output rules (must follow exactly): …
- Surface-form fidelity rules: …
  - Do NOT modernize, normalize, regularize, or paraphrase spelling. …
- Tokenization rules: …
- Allowed tags: …
- Important constraints: …
- Now tag this text: (input)
- Output results in this format: …