Universiteit Leiden

Master Computer Science

Finding Sound in Written Latin Epic

Name: Luuk Nolden
Student ID: 1370898

Date: 13-04-2022

Specialisation: Data Science

1st supervisor: Suzan Verberne
2nd supervisor: Matthew Payne

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This is where you write an abstract that concisely summarizes your thesis. Keep it short. No references here — exceptions do occur.

Contents

1 Introduction

Automated scansion of poetry by computer is a long-standing goal, and a difficult task. Antje Wessels and Matthew Payne (Classics) and Luuk Nolden and Philippe Bors (Computer Science), who have developed a digital platform specifically for Classical fragments (oscc.lucdh.nl), now plan the implementation of a digital tool to automatically scan Latin poetry. They plan to utilize machine learning to bring a more sophisticated approach which plays to the strengths of artificial intelligence.

Why is the study of meter important? The sound of a text is an essential part of its content. Similar to what we experience when we listen to different musical styles, our perception of a text depends on the rhythmical patterns in which this text is presented. Different rhythms lead to different expectations about its content, and across literatures and cultures specific rhythms have always been associated with specific genres. This was as true in the ancient world as in the modern. Just as we expect something very different from hearing the rhythms of slam poetry than those of a limerick or from those of a ballad, so an ancient audience would have instantly recognized a line of the epic writer Ennius from a speech of a slave in Plautine comedy - and the Roman writer Cicero tells us that his fellow Romans were sensitive enough to the meanings of rhythms that they could instantly sense when a writer was subverting expectations through his use of metrical patterns.

Why Latin poetry? Latin poetry features a variety of authors and works which use a wide range of different poetic meters. Actually, rhythm was everywhere in Roman literature, even in prose it was employed as a crucial element: As masters of rhetoric such as Cicero stressed, rhythm, though it should be used subtly, is one of the qualities of good speeches. In Latin poetry different poetic meters - different arrangements of rhythmic patterns - marked out different genres, although the principles of how the sounds of poetry created rhythms remained the same. So Latin poetry provides a large amount of different material written in different poetic meters. Moreover, some of these meters featured simpler rhythms with less metrical variation, others far more complex rhythms with lots of variation. This makes them an excellent set of data for developing a complex scansion tool.

What tools from the digital humanities will you use? Because Latin poetry contains lots of different meters, it is more important to understand the principles by which the sounds of words create the rhythms than any specific patterns that may only work for one author or work. Just like a student needs to learn to apply these principles and to get a feeling for the relationship between these principles and how they work in practice in texts, so too for automatic scansion to succeed, the machine needs to be able to translate between these fundamental principles and specific metrical patterns as they are found in different authors and genres. This is why we will utilize approaches from machine learning, rather than trying to create a model that follows rules which only work for one kind of metrical pattern. Our current plan is to create a model which can train itself using techniques such as Markov chains and SVM, in order that it be able to handle metrical ambiguities and exceptions. In this way the model should even be able to handle a completely new kind of meter.

How will you go about this? Just like we teach a student first how to scan more straightforward metrical lines like the dactylic hexameter before going on to rather complex metrical forms such as found in the odes of Horace or in the polymetrical cantica of Plautus, we are going to train our machine on the corpora of Virgil (written in dactylic hexameter), then introduce a different meter, iambic trimeter, found in the tragedies of Seneca, before finally training the model on the more complex metrical forms of Plautus.

What are some interesting applications? Not all Latin poetry survives to us as complete texts - many lines come to us in the form of fragments, because we find them as quotations in prose texts. Fragments are typically preserved in incomplete metrical units, such as half-lines. They are also more vulnerable to textual corruption, because they were quite likely to be misunderstood or confused with the prose text they were quoted within. One exciting possibility is that the model will help us to better identify the metrical pattern of a corrupt line.

Even more exciting is the possibility of using the model to detect new fragments. There are many prose texts which scholars believe may embed verse lines, but it is difficult and time-consuming for an individual to look through all of them. Applying the model to these texts could flag passages which seem to embed such lines.

What relevance does this have beyond Latin poetry? A model trained to successfully recognize and extrapolate rhythmic information from texts written in Latin could likely be adapted to work in any language. The adaptation might involve changing the model from analyzing quantity (as in Latin) to analyzing stress (as, for instance, in most Dutch or English poetry), but, while a rules-based approach would require starting from scratch, the machine learning approach could accommodate this.

Who will benefit from this project? Any student of Latin poetry, at whatever level, who is trying to get a better understanding of metre and its relevance to the content of poetry will gain from applying the model to the texts they are learning about. Rather than seeing scansion as a mechanistic process of applying rules, by experimenting with the scansion tool they will hopefully gain greater insight into the possibilities and complexities of different metres.

As we said at the beginning, the sound of a text is an essential part of its content. By applying machine learning techniques to this problem of metre through the large and multi-faceted body of data provided by Latin poetry we aim to test their usefulness and extensibility. And we hope that any researchers in the humanities who are working on topics where sound is connected with meaning will benefit from our exploration of developing a model through machine learning.

2 Skeleton

Can a neural network learn to scan Hexameters?

1. Approach using HMM 2. Approach using CRF 3. Approach using LSTM 4. Approach using ELECTRA
What is the effect of different genres For Hexameter, two different genres: saturae and epic. No noticeable difference detected.
Clustering texts: agglomerative clustering: create tree from Kestemont
What is the effect of different time periods No difference detected. Virgil works fine on Iuvenal.
Notes: normalise results. same corpus size ovid compared to same size virgil
Flair What is the effect of different word representations? Strings, hashes, syllable embeddings, char/syll/word embeddings Different context windows (5), dimensions (100-300), no mincount fasttext (autovocab), glove, word2vec cluster these vectors using PCA: are clusters logial and relevant to our task? syntactical and semantical?
How well do these networks generalise to other meters? Elegiac, Iambic trimeter, anapest, trochaic tetrameter, polymetrum, hendycasyllabi

2.1 Conditional Random Fields

Notes:

1. Passing only syllables gave lower accuracy (between 1 and 5%)
2. More features like diphthongs or consonants did not improve accuracy
3. No time era difference detected. Virgil works fine on Iuvenal.
4. For Hexameter, two different genres: saturae and epic. No noticeable difference detected.
5. Kestemont did detect genres

```python
features = {
    'bias': 1.0,
    '0:syllable' : word, # just take the entire syllable
    '0:last_3_char': word[-2:], # Take last 2 characters
    '0:last_2_char': word[-1:], # Take last 1 characters
}
# Check if we are at the beginning of the sentence
if i == 0:
    features['BOS'] = True # This should always be long
# Gather features from the previous word
if i > 0:
    previous_word = sent[i-1][0]
    features.update({
        '-1:word': previous_word,
        '-1:last_1_char': previous_word[-1:],
        '-1:last_2_char': previous_word[-2:],
    })
# Gather features from the next word
if i < len(sent)-1:
    next_word = sent[i+1][0]
    features.update({
        '+1:word': next_word,
        '+1:first_1_char': next_word[:1],
        '+1:first_2_char': next_word[:2],
    })
else:
    features['EOS'] = True # This will be an anceps
```

Listing 1: Python example

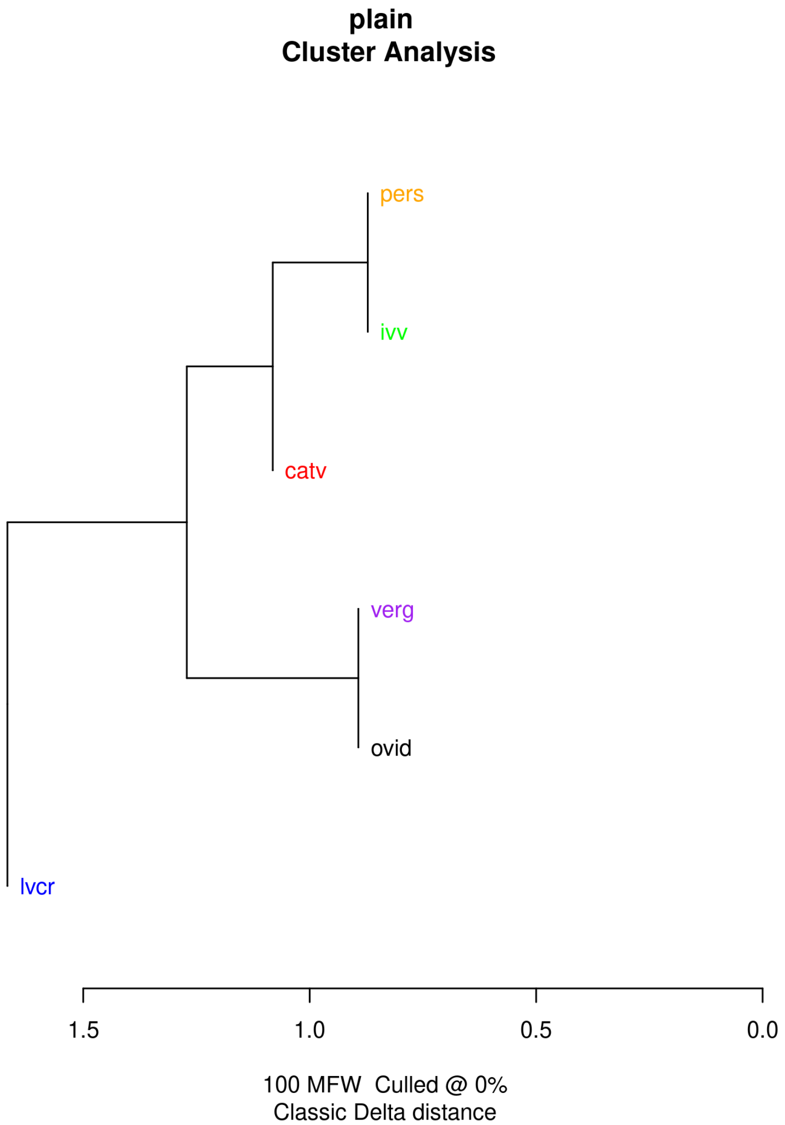
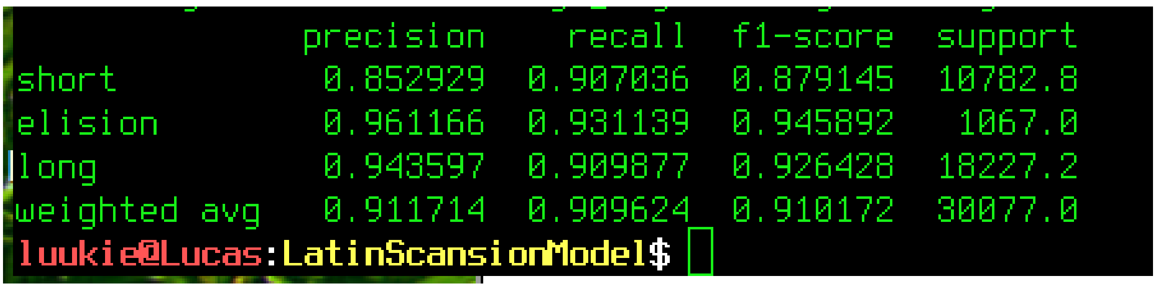


Figure 1: Hexameter texts clustering.



	precision	recall	f1-score	support
short	0.852929	0.907036	0.879145	10782.8
elision	0.961166	0.931139	0.945892	1067.0
long	0.943597	0.909877	0.926428	18227.2
weighted avg	0.911714	0.909624	0.910172	30077.0

luukie@Lucas:LatinScansionModel$

Figure 2: CRF confusion matrix for Virgil's Aeneid.

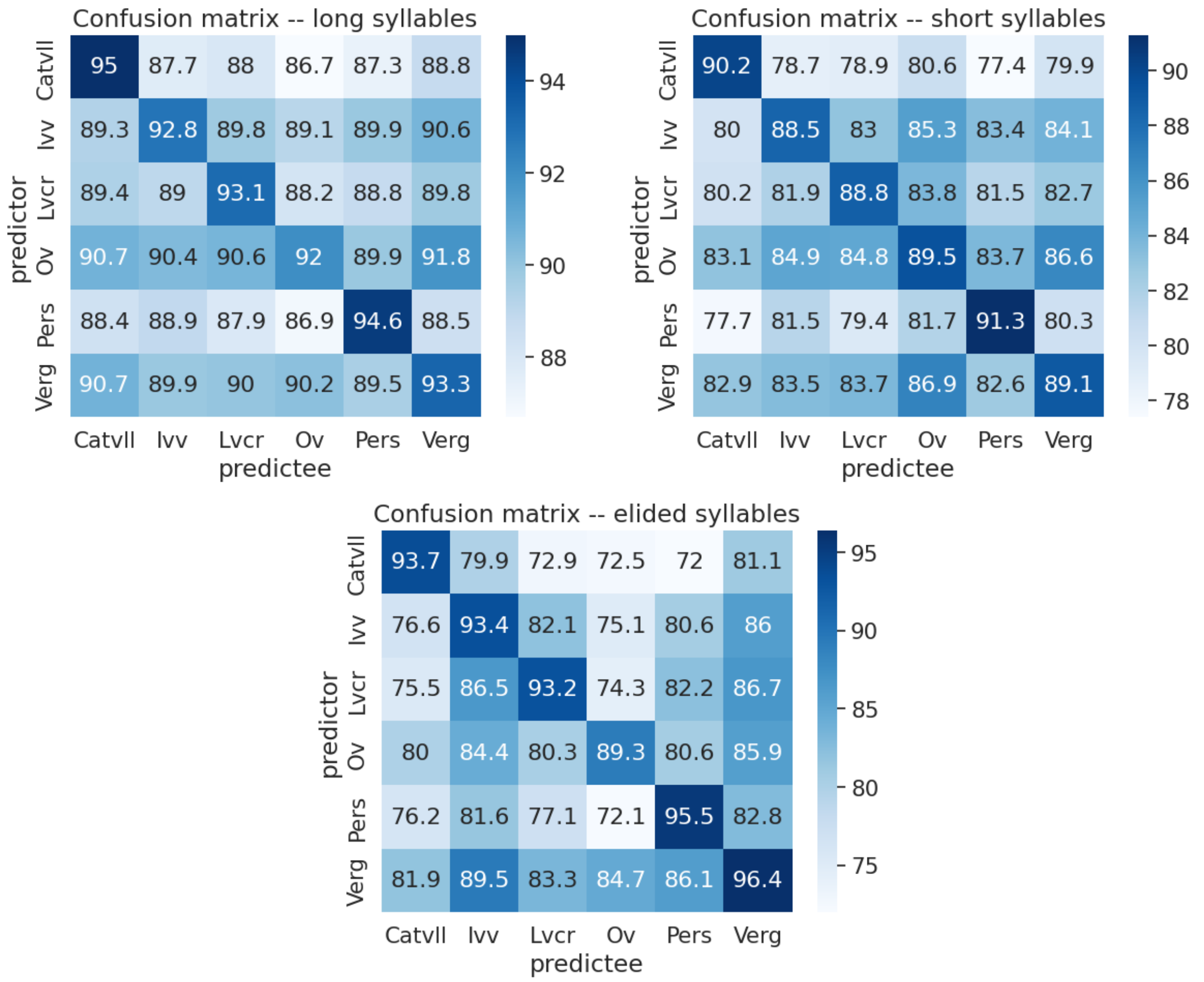


Figure 3: CRF label accuracy for hexameter texts.

2.2 LSTM

1. Syllables encoded with integers. Five labels: long, short, elision, space and padding.
2. 25 epochs used, bi-directional LSTM, very simple.
3. LSTM has extremely high accuracy within author if text is larger than 1000 lines.
4. Transferability between authors is better than CRF.
5. As with CRF, texts have to be larger than 1000 lines to achieve a nice accuracy.

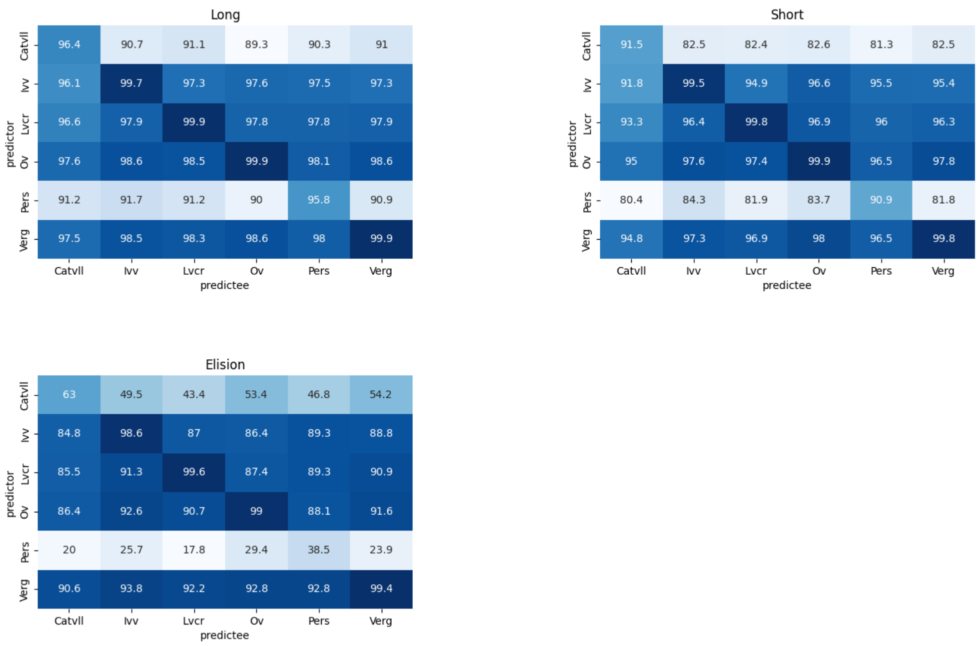


Figure 4: LSTM label accuracy for hexameter texts.

	precision	recall	f1-score	support
long	0.9901	0.9929	0.9915	18207
short	0.9863	0.9823	0.9843	10822
elision	0.9732	0.9661	0.9696	1090
micro avg	0.9881	0.9881	0.9881	30119
macro avg	0.9832	0.9804	0.9818	30119
weighted avg	0.9881	0.9881	0.9881	30119

Figure 5: LSTM confusion matrix for Virgil's Aeneid.

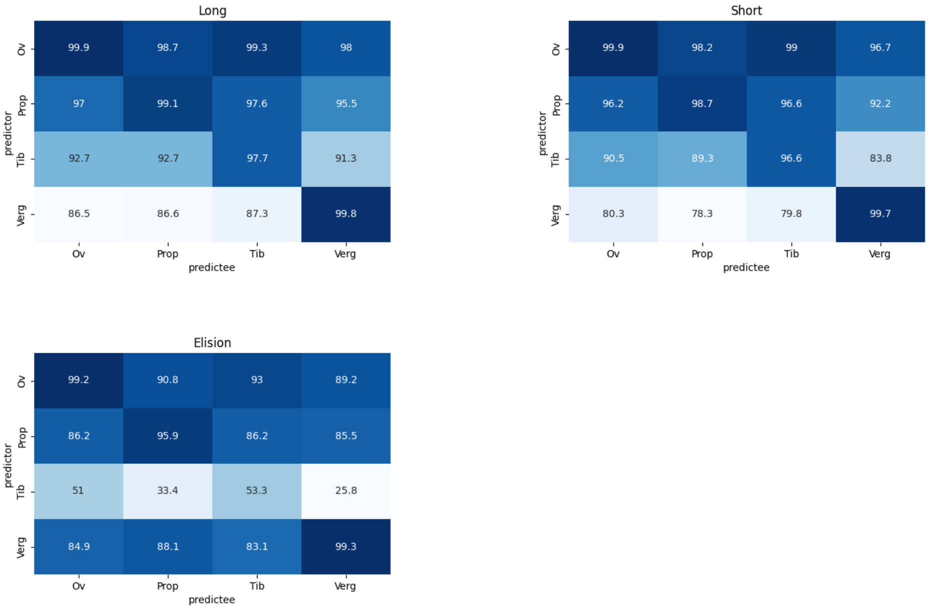


Figure 6: LSTM label accuracy for elegiac texts.

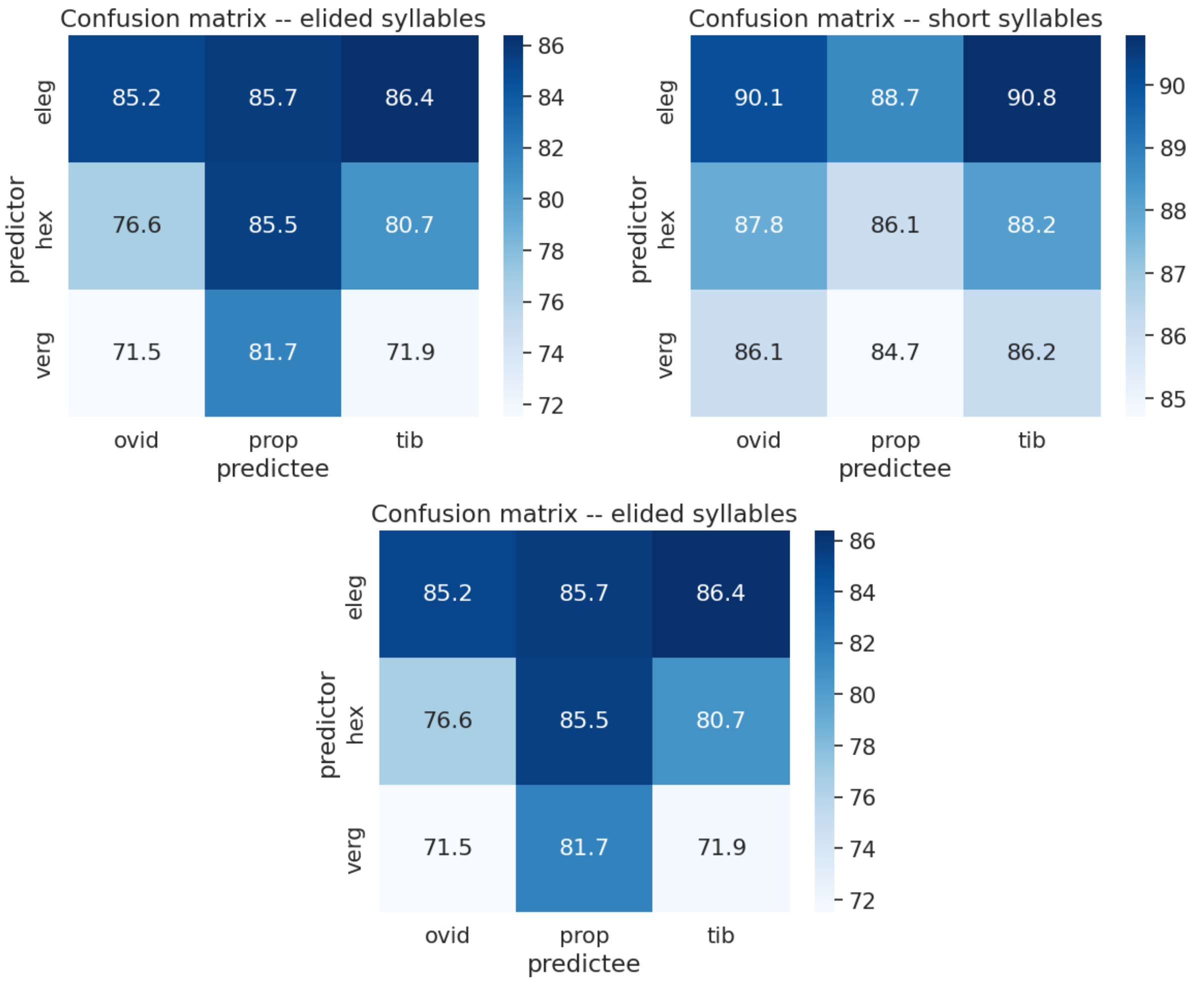


Figure 7: LSTM label accuracy when combining hexameter and elegiac texts.

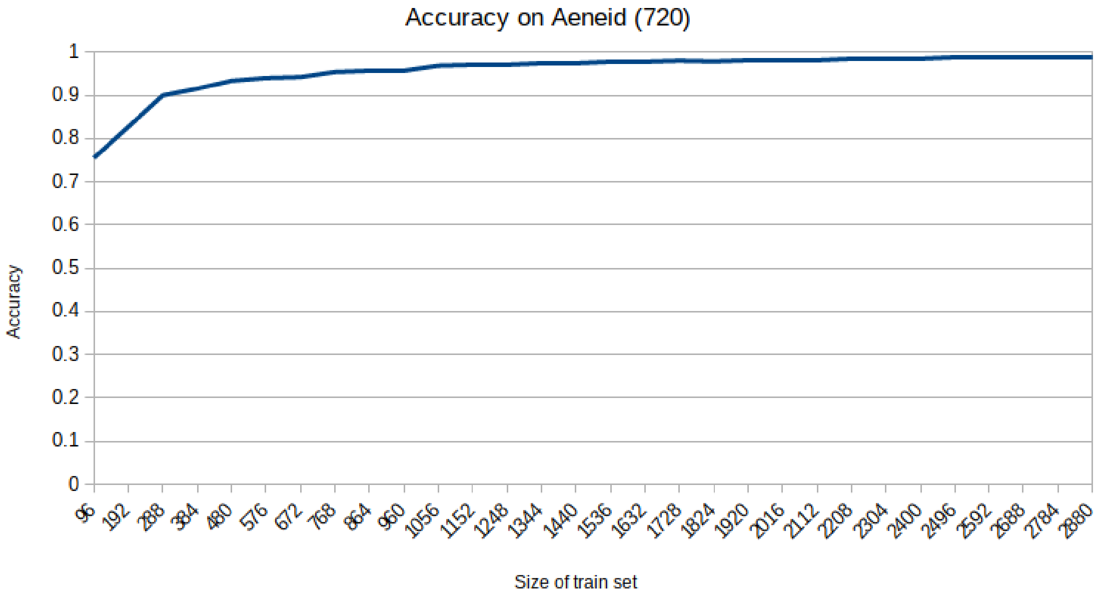


Figure 8: LSTM minimum train size for acceptable accuracy.

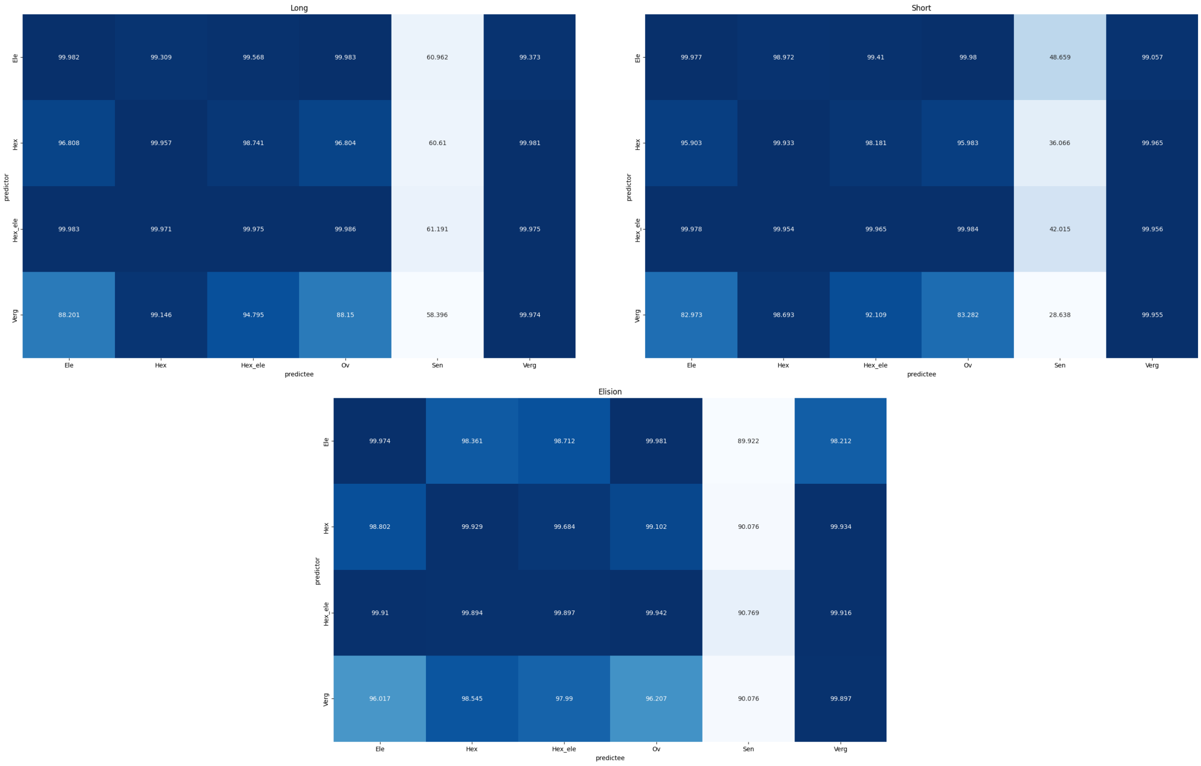


Figure 9: LSTM label accuracy when combining hexameter and elegiac texts (trimeter).

3 Background

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

4 Data and Implementation

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5 Analysis and Results

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

# of class	label	Trained Data		New Data	
		Prec.	Rec.	Prec.	Rec.
3	long	60.6	100.0	??	??
	short	0.0	0.0	??	??
	elision	0.0	0.0	??	??
Average		36.7	60.6	??	??

Table 1: Scores for the base line (predict everything as long)

# of class	label	Trained Data		New Data	
		Prec.	Rec.	Prec.	Rec.
3	long	69.7	94.1	??	??
	short	73.2	37.2	??	??
	elision	100.0	0.0	??	??
Average		81.0	72.0	??	??

Table 2: Scores for the hidden markov model

# of class	label	Trained Data		New Data	
		Prec.	Rec.	Prec.	Rec.
3	long	94.4	91.0	??	??
	short	85.3	91.0	??	??
	elision	96.1	93.1	??	??
Average		91.2	91.0	??	??

Table 3: Scores for the crf sequence labeler

6 CRF Sequence Labeling

Feature based on position from the end (to find fifth and sixth foot): using position from last syllable index is overfitting on hexameters. Would we call a foot a named entity?

7 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

8 Conclusion

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

References

Aland, K. *Synopsis Quattuor Evangeliorum*. Stuttgart, 1985.

Argyle, A.W. "Evidence for the View that St. Luke used St. Matthew's Gospel". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1964, pp. 371–379.

Augustine, St. "The Harmony of the Gospels". In: *Nicene and Post-Nicene Fathers: Series I/Volume VI*. Ed. by P. Schaff. Oxford, 1885.

Barret, C.K. "Q: A Reexamination". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1942-1943, pp. 259–268.

Bellinzoni, A.J. *The Two-Source Hypothesis*. Georgia, US, 1985.

Bradby, E.L. "In Defence Of Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1956, pp. 287–293.

Carlson, S.C. *The Two Source Hypothesis*. 1999. URL: `https://web.archive.org/web/20080515022301/http://www.mindspring.com/~scarlson/synopt/2sh/index.htm` (visited on 02/06/2020).

— *The Two Source Hypothesis*. 1999. URL: `https://web.archive.org/web/20080515022301/http://www.mindspring.com/~scarlson/synopt/2sh/index.htm` (visited on 07/10/2020).

Dalen-Oskam, K. van and J. van Zundert. "De list van het lexicon. Auteursonderscheiding met behulp van computer-ondersteunde woordenschatanalyse". In: *Nederlandse Letterkunde.* 10 (2005), pp. 212–231.

Damerau, F.J. "The use of function word frequencies as indicators of style". In: *Computers and the Humanities* 9 (1975), p. 271.

Deneire, T. "Filelfo, Cicero and epistolary style : a computational study". In: *Brill's studies in intellectual history* 289 (2018), pp. 239–270.

Downing, F.G. "Towards the Rehabilitation of Q". In: *NTS* 11 (1964-1965), pp. 169–181.

Dungan, D.L. "Critique of the Q Hypothesis". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1970, pp. 427–433.

Eder, M., J. Rybicki, and M. Kestemont. *'Stylo': a package for stylometric analysis*. 2017.

Eichhorn, J.G. "Über die drei ersten Evangelisten". In: *Allgemeine Bibliothek der biblischen Literatur* 5 (1794), p. 766.

Engelbrecht, J. "Challenging the two-source hypothesis: how successful are the commentaries?" In: *Neotestamentica* 30(1) (1996), pp. 89–101.

Farmer, W.R. "A Fresh Approach to Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1975, pp. 397–408.

Farrer, A.M. "On Dispensing with Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1985, pp. 321–356.

Fitzmyer, J.A. "Luke's Use Of Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1970, pp. 245–257.

Goodacre, M. *Ten Reasons to Question Q*. 1997. URL: `https://web.archive.org/web/20081015182345/http://ntgateway.com/Q/ten.htm` (visited on 02/06/2020).

— *Ten Reasons to Question Q*. 1997. URL: `https://web.archive.org/web/20081015182345/http://ntgateway.com/Q/ten.htm` (visited on 07/11/2020).

Griesbach, J.J. "A Demonstration that Mark was Written after Matthew and Luke". In: *J.J. Griesbach: Synoptic and Text - Critical Studies 1776–1976*. Ed. by B. Orchard and T.R.W. Longstaff. Vol. 34. 1978, pp. 74–102.

Grieve, J.W. *Quantitative Authorship Attribution: A History and an Evaluation of Techniques*. 2002.

Hawkins, J.C. "Three Limitations to St. Luke's Use of St. Mark's Gospel: The Disuse of the Markan Source in St. Luke ix. 51-xviii. 13". In: *Oxford Studies in the Synoptic Gospels*. Ed. by W. Sanday. 2nd ed. 1911, pp. 29–59.

Holmes, D. "The Evolution of Stylometry in Humanities Scholarship". In: *Literary and Linguistic Computing* 13 (1998), pp. 111–117.

Honoré, A.M. "A Statistical Study of the Synoptic Problem". In: *Novum Testamentum* 10 (1986), pp. 95–147.

Inglis, D. *Tricky NT Textual Issues*. URL: `https://sites.google.com/site/inglisonmarcion/Home/the-synoptic-problem/the-making-of-the-double-tradition` (visited on 07/06/2020).

Kestemont, M., S. Moens, and J. Deploige. "Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux". In: *DSH* 30 (2014), pp. 199–224.

Kloppenborg, J.S. *Excavating Q: The History and Setting of the Sayings Gospel*. Minneapolis, US, 2000.

Kümmel, W.G. "In Support of Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1973, pp. 227–243.

Lummis, E.W. *How Was Luke Written: Considerations Affecting the Two-Document Theory with Special Reference to the Phenomenon of Order in the Non-Markan Matter Common to Matthew and Luke*. Cambridge, 1915.

Marxsen, W. *Introduction to the New Testament: An Approach to its Problems*. Michigan, US, 1968.

Mascol, C. (a.k.a. W.B. Smith). "Curves of Pauline and Pseudo-Pauline Style I&II". In: *Unitarian Review* 30 (1888), 452–460 and 539–546.

Montserrat, J. *Two-Source Hypothesis*. 2005. URL: `http://www.hypotyposeis.org/synoptic-problem/2004/09/two-source-hypothesis.html` (visited on 02/06/2020).

Perseus, Digital Library. *Perseus Vocabulary Tool*. 2007. URL: `http://www.perseus.tufts.edu/hopper/help/vocab` (visited on 05/18/2020).

Petrie, S. "'Q' is Only What You Make It". In: *Novum Testamentum* 3 (1959), pp. 28–33.

Sanders, E.P. "The Argument from Order and the Relationship between Matthew and Luke". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1968, pp. 409–426.

— *The Tendencies of the Synoptic Tradition*. Cambridge, 1969.

Simpson, R.T. "The Major Agreements of Matthew and Luke Against Mark". In: *NTS* 12 (1965-1966), pp. 273–284.

Streeter, B.H. "The Document Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1985, pp. 221–225.

— *The Four Gospels: A Study of Origins, Treating of the Manuscript Tradition, Sources, Authorship, & Dates*. Virginia, US, 1924.

Styler, G.M. "The Priority of Mark". In: *The Birth of the New Testament*. Ed. by C.F.D. Moule. 1962, pp. 223–232.

Taylor, V. "The Original Order of Q". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1959, pp. 295–317.

Tyson, J.B. "The Two-Source Hypothesis: A Critical Appraisal". In: *The Two-Source Hypothesis: A Critical Appraisal*. Ed. by A.J. Bellinzoni. 1985, pp. 437–452.

Watson, F. “Q as Hypothesis: A Study in Methodology”. In: *New Testament Studies* 55 (2009), pp. 397–415.

Westcott, B.F. and F.J.A. Hort. *The New Testament in the Original Greek*. Cambridge, 1885.

Woods, F.H. “The Origin and Mutual Relations of the Synoptic Gospels”. In: *Studia Bibblica et Ecclesiastica* II (1886), pp. 59–104.