Style in Eight Syllables

Metric Annotation and Stylometry of Chrétien de Troyes and Contemporaries

Florian Cafiero Philippe Chaumet-Riffaud **Damien Conceicao†**
Ulysse Godreau Émilie Guidi Alexandre Lionnet **Théo Moins†**
Pierre-Alexandre Nistor Benedetta Salvati **Jean-Baptiste Camps†**

École nationale des chartes, Université PSL

DH Benelux 2025
Amsterdam, 5 June 2025, *Session 4a*

European Research Council
Established by the European Commission

Prosodic annotation for the attribution of Medieval texts

Known to increase attribution performance of verse (and even prose) texts

cf. Plecháč and Birnbaum, 2023.

More stable authorial features for medieval texts?

(*ChevLyon*, v. 3501)

H **Chiet** del **fuer**re si li a**point**e

V **Chiet** dou **fuer**re et si a**point**e

F **Cait** a **te**re si li a**point**e

R **Kiet** del **fuer**re si que li **point**e

- Graphemic and lexical variation (incl. FW)
- but identical prosodic pattern (SwWswwwS) (slightly more robust than rhyme word)

Case study: the attribution of the works of Chrétien de Troyes

The case in 30 seconds

Chrétien de Troyes (c. 1130 – † b. 1180 and 1190), most canonical 12th cent. French writer, and 'introducer' of the Grail;

Collaborative authorship *Lancelot* (with. Godefroi de Leigny)

Dubious attributions Philomena (*Crestiiens li gois*), Guillaume d'Angleterre (*Crestiiens*); Continuations to Perceval;

Case study: the attribution of the works of Chrétien de Troyes

Last year's results

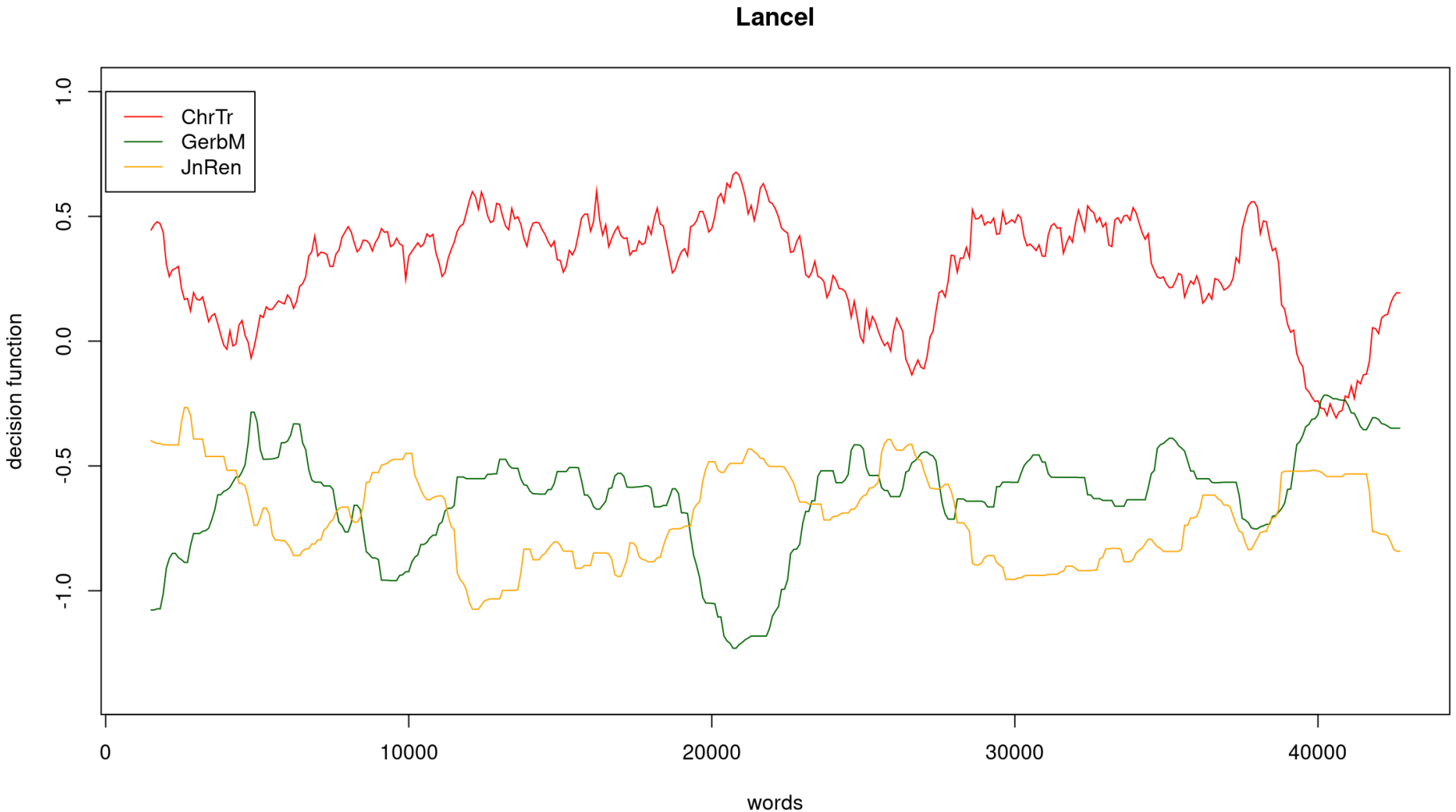
Follow-up to: Camps et al., *The Authorship of the Works of Chrétien de Troyes: A Stylometric Examination*, **DHBenelux 2024**.

Main conclusions:

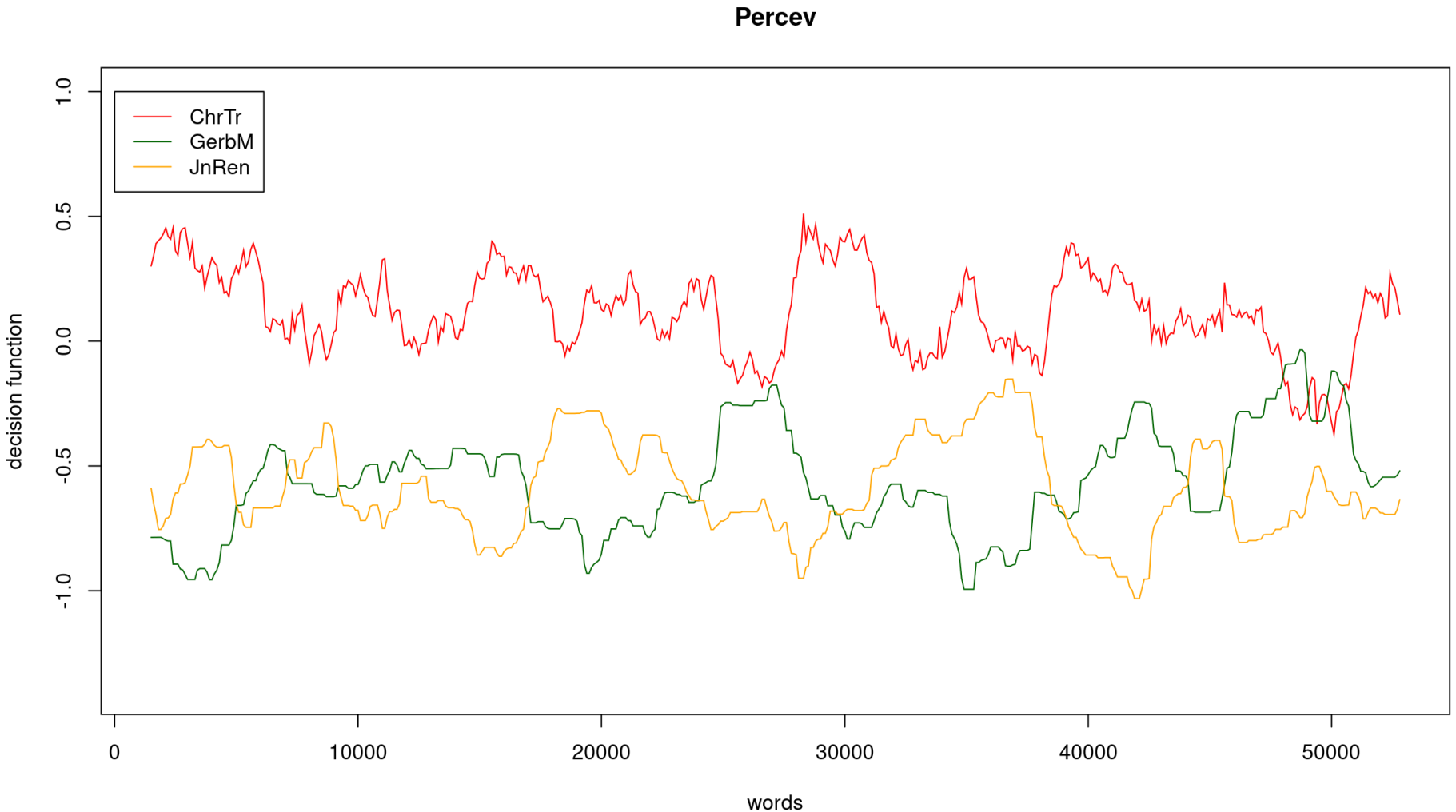
- Lancelot: drop in Chrétien's style around v. 6150, consistent with intervention of Godefroi de Leigny;
- Perceval: new drop towards the end (possible unknown collaborative authorship or artefact?)
- Guillaume d'Angleterre and Philomena: mixed results.

Question: Could rhythm-based patterns perform better than (and/or improve the attribution accuracy of) function words?

2024 Results: Lancelot



2024 Results: Perceval



Outline

Verse types and Metronome annotation pattern

The narrative octosyllabic couplet

- 8 syllable long
- rhymed two by two
- masculine or feminine rhyme
- no standard caesura

Metronome annotation (Nagy et al., 2024)

w weak syllable

S stressed syllable

. word break

Verse	Annotation
et li chevax tant s'esforça	w.w.wS.w.wwS
Et a la pierre de l'anel,	w.w.w.Sw.w.wS

Word breaks were removed for analysis, as they decrease SVM performance

Automated annotation: LLMs vs rule-based

Two approaches

LLMs chain-of-thought; prompting with examples, and a set of expert rules;

Rule based deep-learning for linguistic annotation (POS, lemmas) + rule-based syllabification and stress annotation (using phonological knowledge).

Steps

1. linguistic annotation (rule based only): POS tagging to find function words;
2. syllabification;
3. identification of stressed/unstressed syllables.

Syllabication by rules

Each verse is processed letter by letter using the following rules:

- Basic Rules:
  - Automatic addition of consonants to the current syllable.
  - Addition of vowels followed by the termination of the vowel.
- Complex Rules:
  - If the last syllable of a word contains only consonants, it is attached to the preceding syllable.
  - If the analyzed vowel and the following vowel form a diphthong that does not end in 'e', both are added to the current syllable.
  - If the analyzed vowel and the following vowel form a diphthong that ends in 'e', it is checked whether the 'e' is at the end of the word; if so, it is separated into its own syllable.
  - Unknown characters are removed.
  - Apostrophes are treated as consonants

Results of annotation

- LLMs performed well on syllabification (≥ 0.95), but poorly on stress identification (0.7827 with Claude 3.5).
- Rule-based approach obtained the best results (accuracy 0.9787).

(Levenshtein ratio: $1 - (\text{Levenshtein distance}_{i,j}/\text{len}_i + \text{len}_j)$)
Code: `https://github.com/PoidsPlume/AnoChre`.

Outline

Setup

Features

- function words;
- prosodic:
  - n-grams of syllables;
  - n-grams of verses.

Training

- SVM Linear classifier
- 375 verses samples ($\approx$ 3000 words)
- group-k-fold cross-validation
- downsampling (for imbalance)
- culling at different %ages

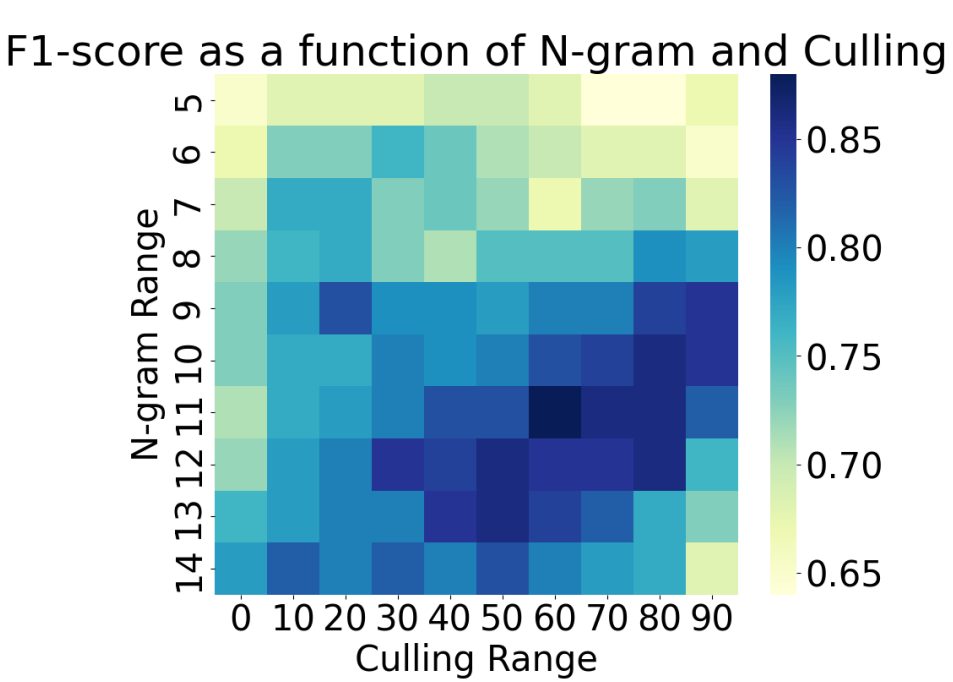
Benchmark results

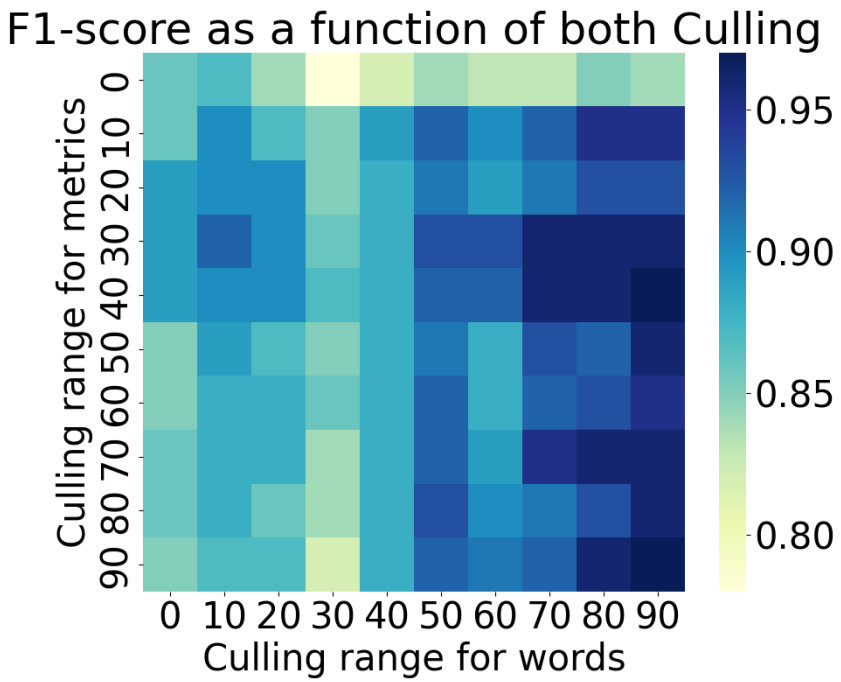
Feats	**n**	**culling**	**F1 score**
funct. words	1	0.1	*0.86*
met. syll.	11	0.6	**0.88**
met. lines	1	0.3	0.75
funct. words	1	0.6	0.95
+ met. syll.	9	0.7	
funct. words	1	0.4	**0.97**
+ met. lines.	1	0.9	

Summary

- taken alone, prosodic features perform comparably to FW;
- in combination, can improve substantially the scores.

Culling and n-gram size (prosodic feats)

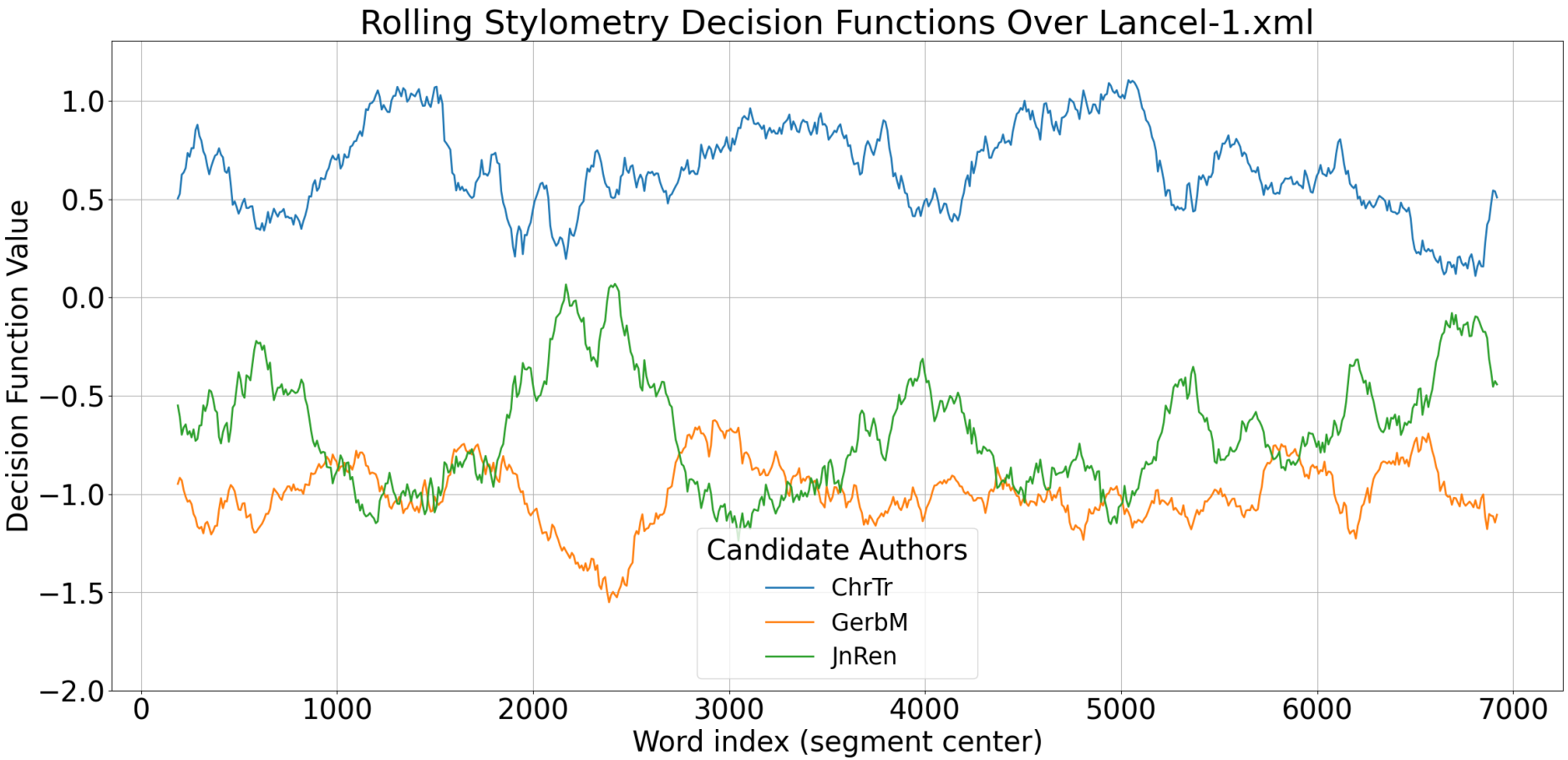




Syllables n-grams length versus culling (left) and culling of function words versus of full verses (right).

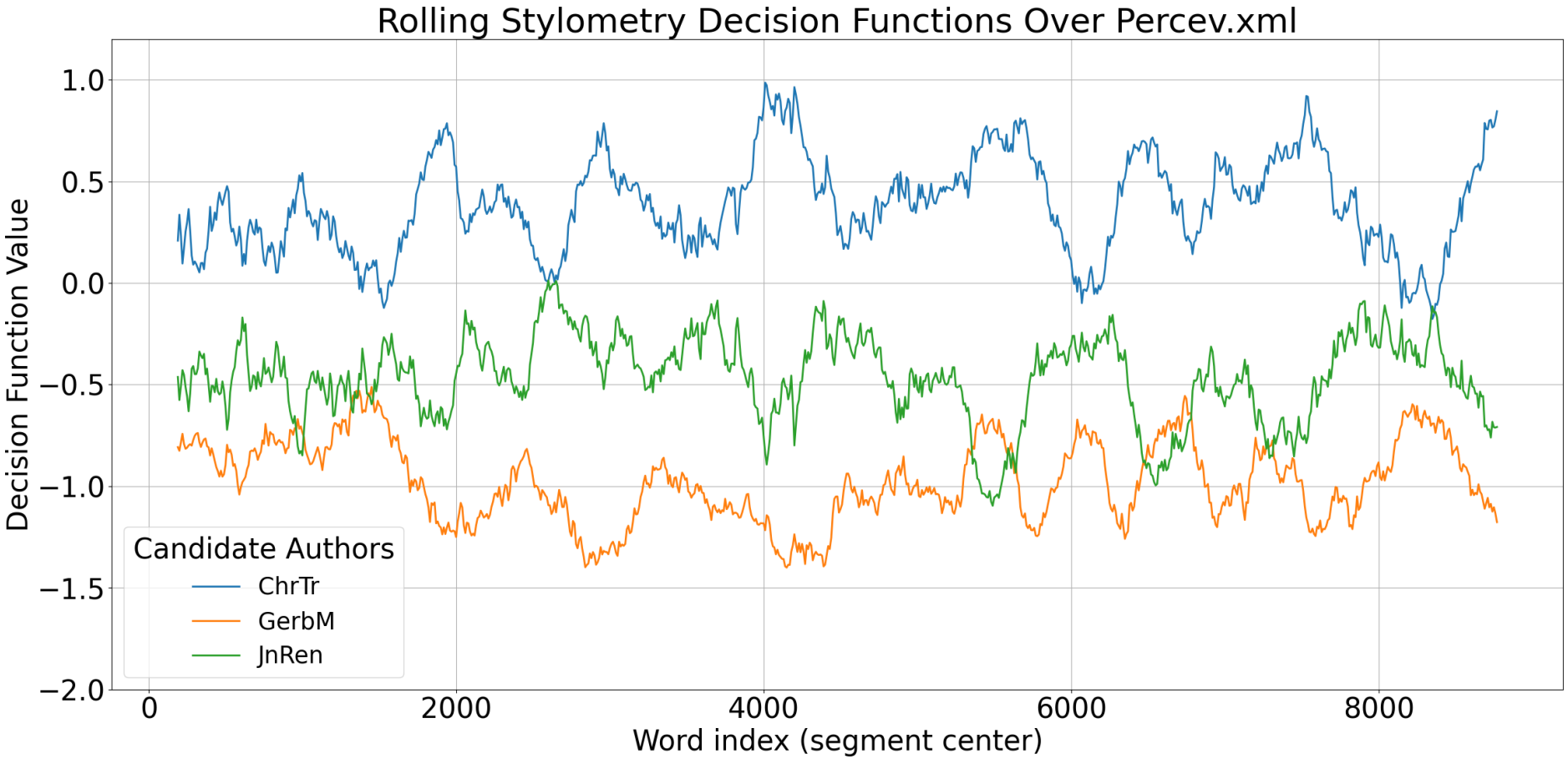
Outline

Lancelot

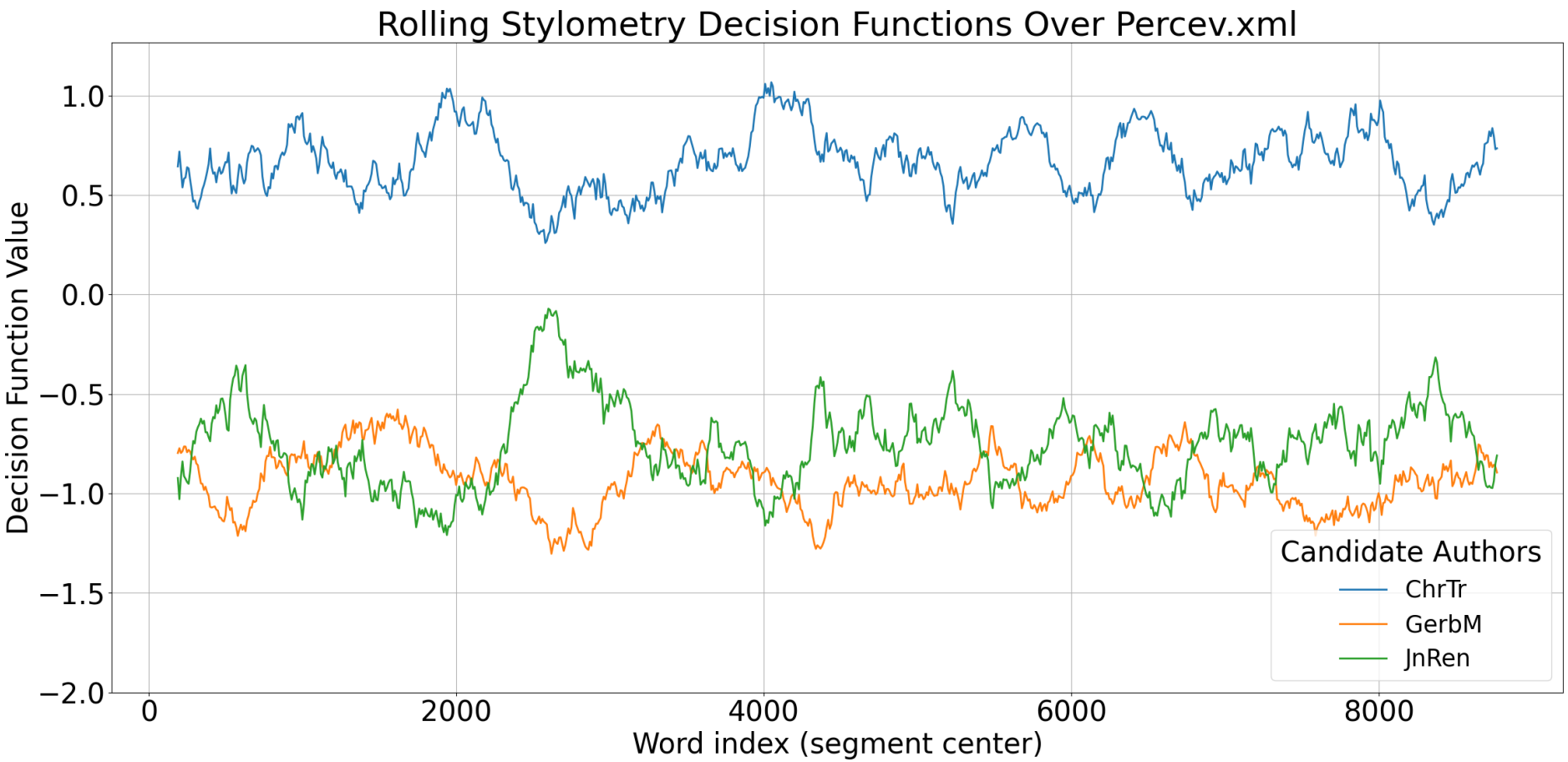


Same drop around v. 6150.

Perceval (FW, met. syll.)

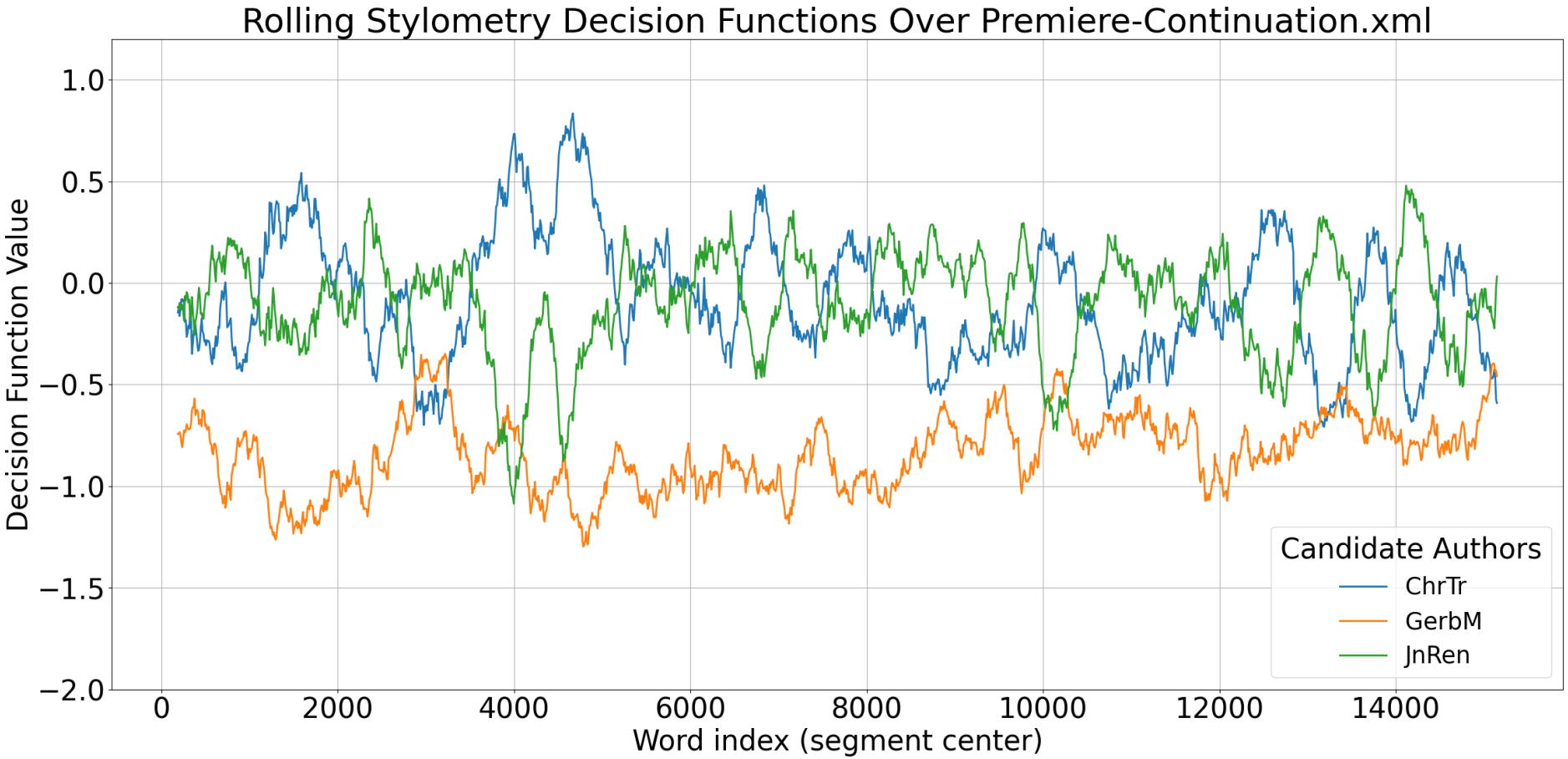


Perceval (FW, met. line)

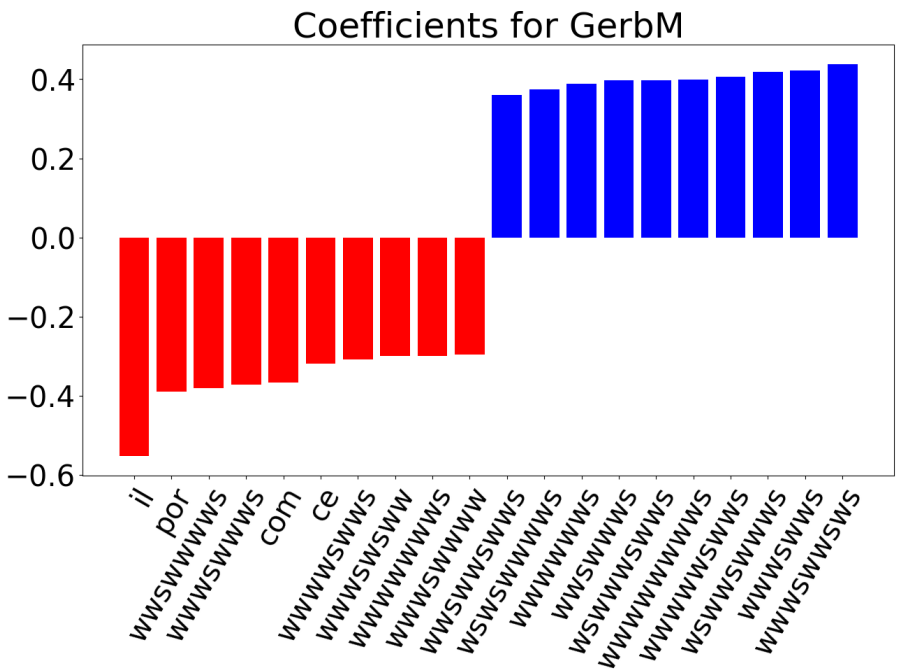


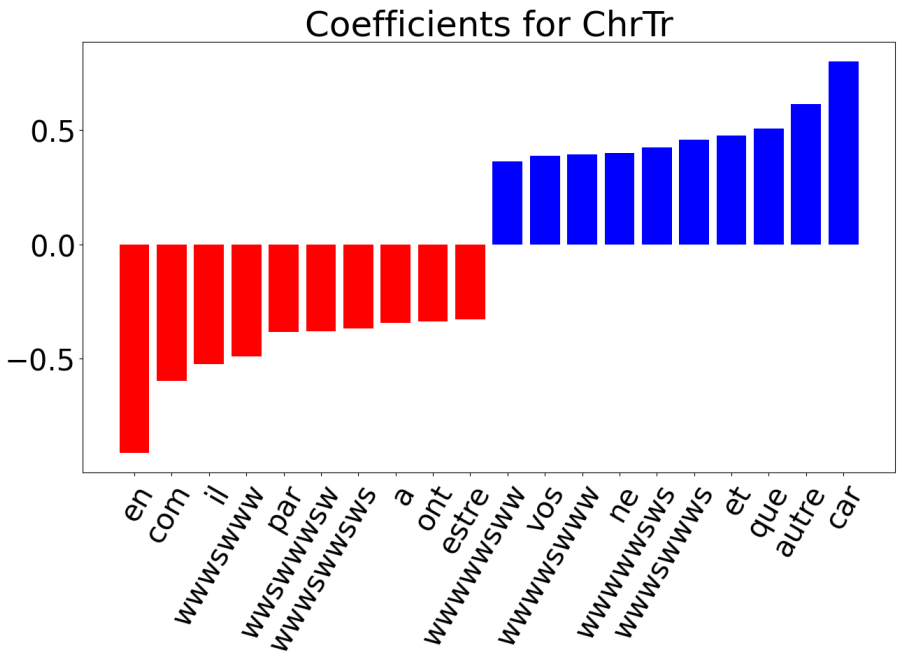
Last year's blip towards the end in Chrétien's style disappeared. . .

First Continuation



Coefficients





Discussion

Main conclusions

1. Prosodic features function reasonably well in isolation, and can improve substantially the baseline results (function words) for Old French verses;
2. features most efficient in isolation are not necessarily the most useful in combination (need to bring additional information);
3. collaborative status of Lancelot again confirmed?
4. status of Perceval: blip disappears with best features;
5. (highly culled) prosodic features seem to give an avantage to Chrétien (more distinct style?).

More research needed

- surprising behaviour of culling at different levels;
- authorship verification approach needed with the new features (e.g. impostors).

Thanks

To the students!

This paper results from the collective work, made by the students of the MA Computational Humanities (Université PSL), in the Seminar of *Computational Philology*, under the supervision of Jean-Baptiste Camps, Florian Cafiero & Benedetta Salvati (year 2024-2025)

To the European Research Council

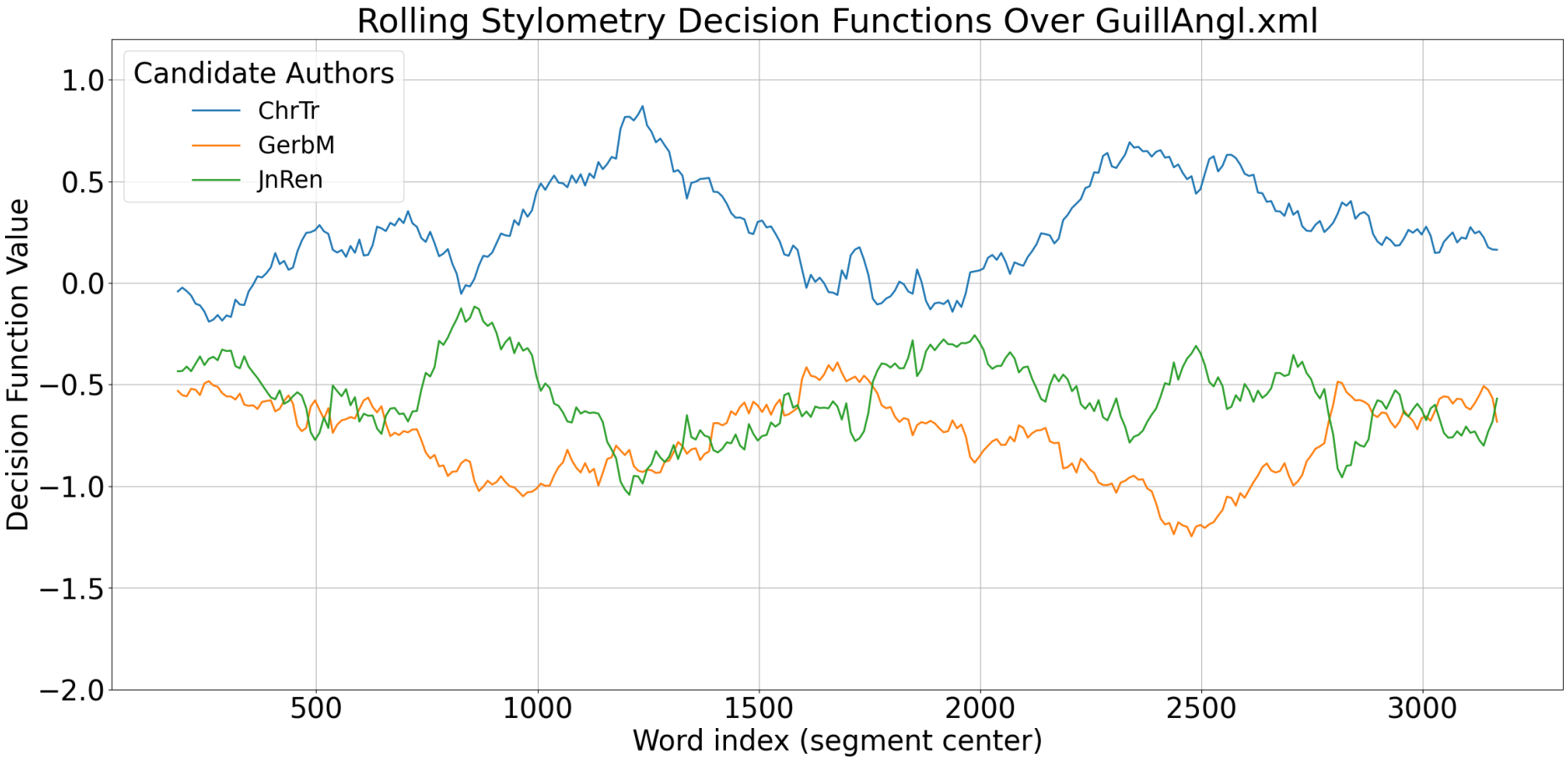
It benefited from funding from project The Lost Manuscripts of Medieval Europe: Modelling the Transmission of Texts (LostMA; ERC StG, 101117408).

To the École des chartes, PSL

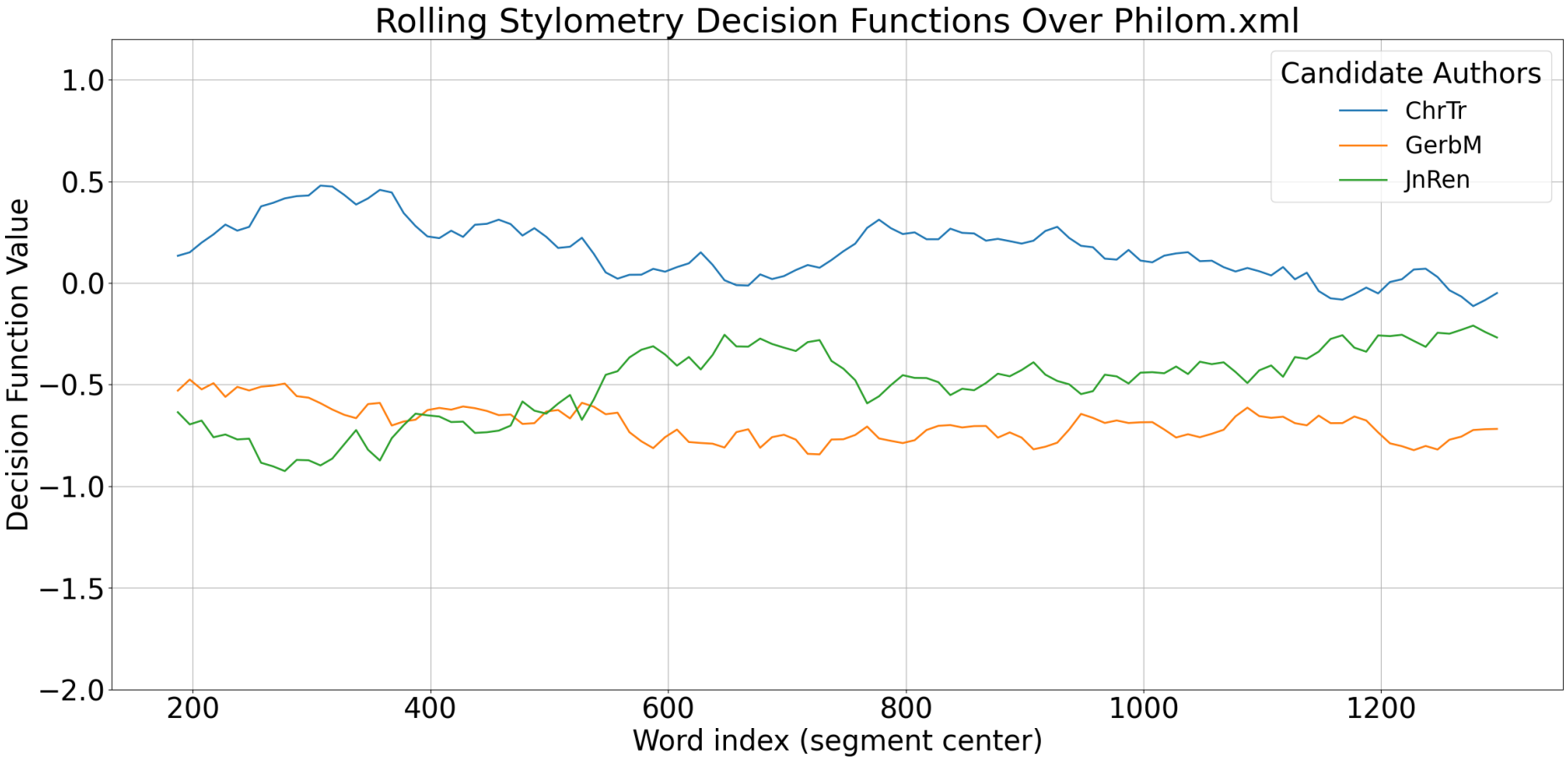
Student participation was funded by the École nationale des chartes, PSL.

Outline

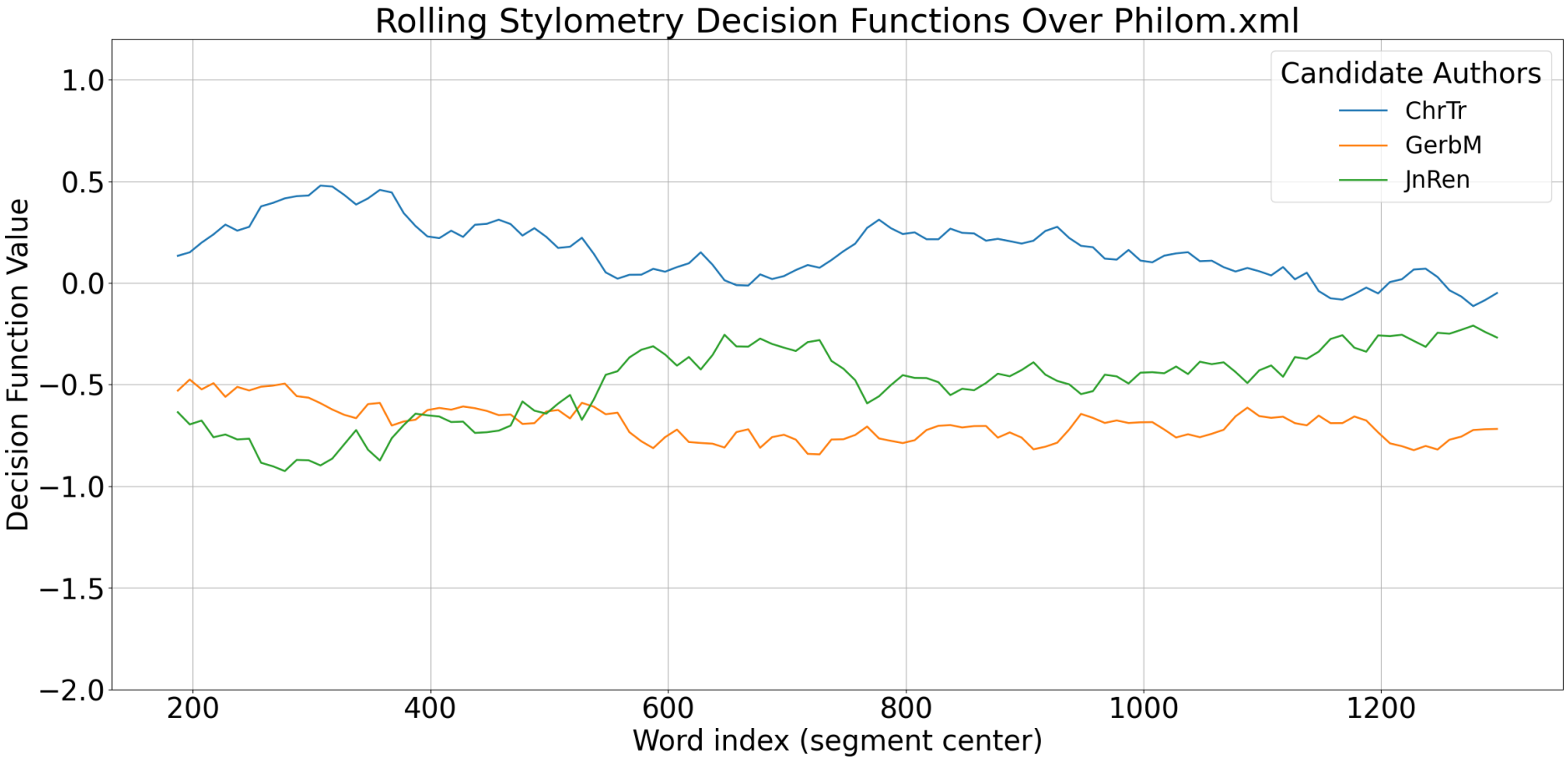
Guillaume d'Angleterre



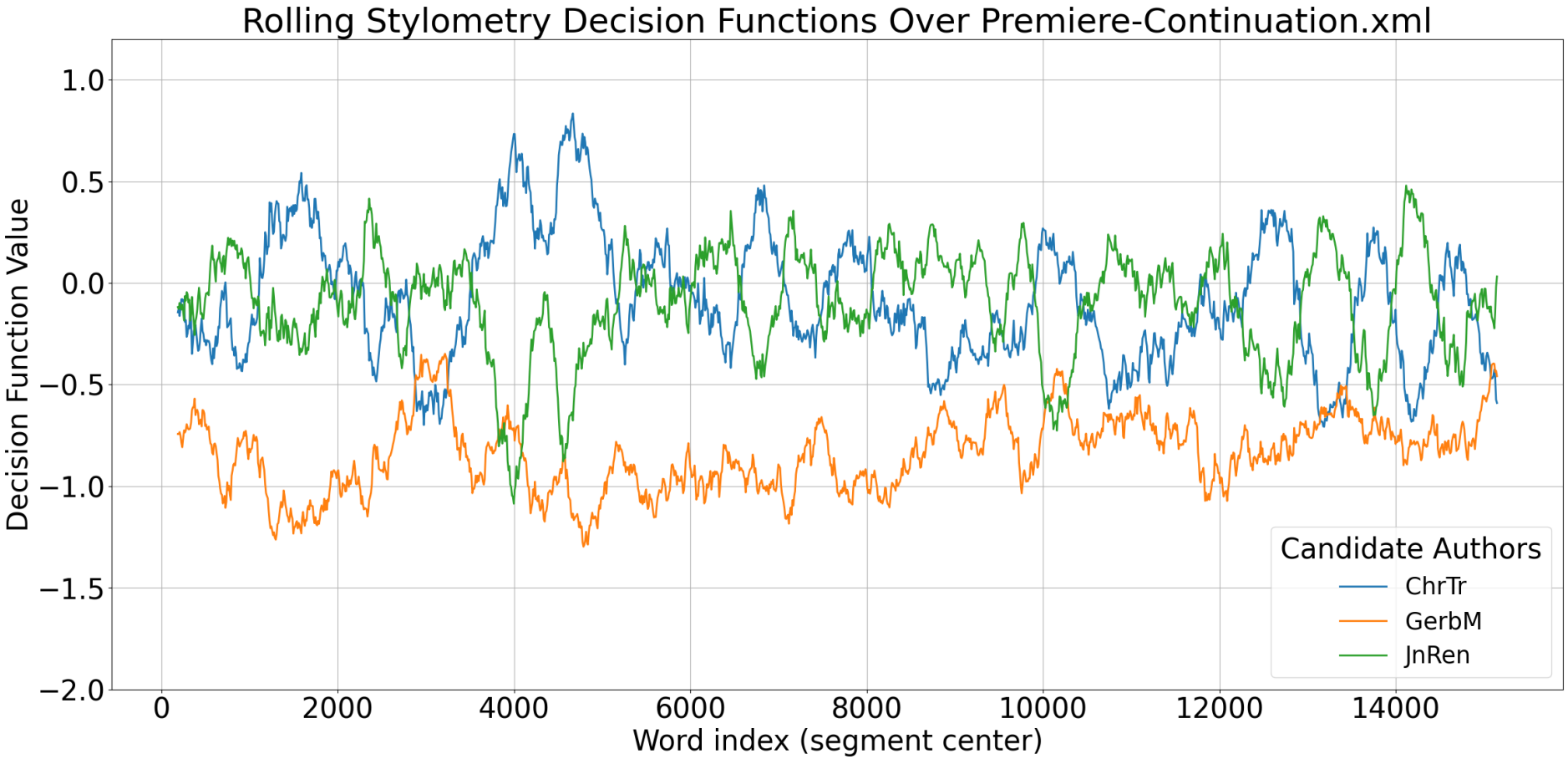
Philomena


Rolling Stylometry Decision Functions Over Philom.xml
Candidate Authors
ChrTr
GerbM
JnRen
Decision Function Value
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
200
400
600
800
1000
1200
Word index (segment center)

Philomena

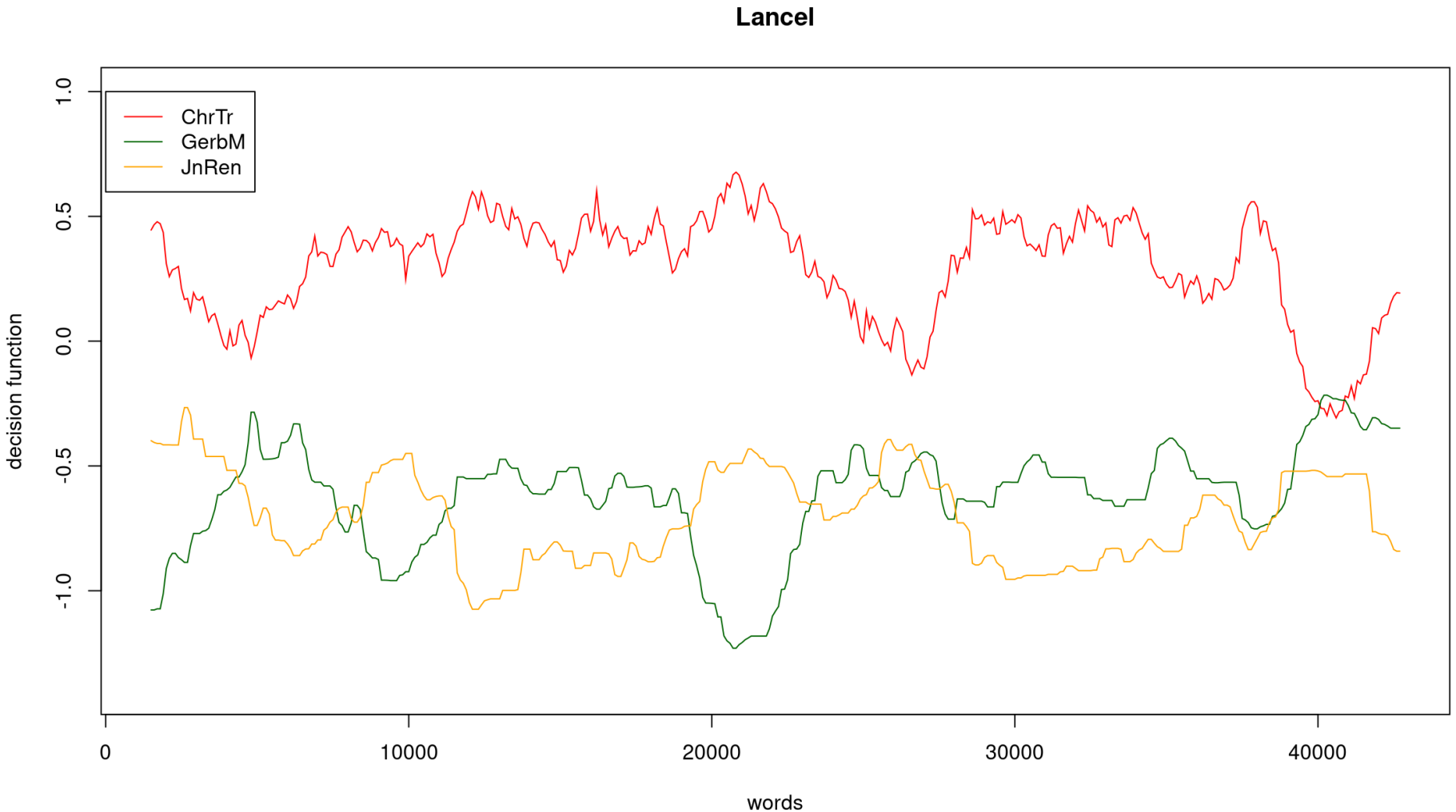

Rolling Stylometry Decision Functions Over Philom.xml
Candidate Authors
ChrTr
GerbM
JnRen
Decision Function Value
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
200
400
600
800
1000
1200
Word index (segment center)

First Continuation

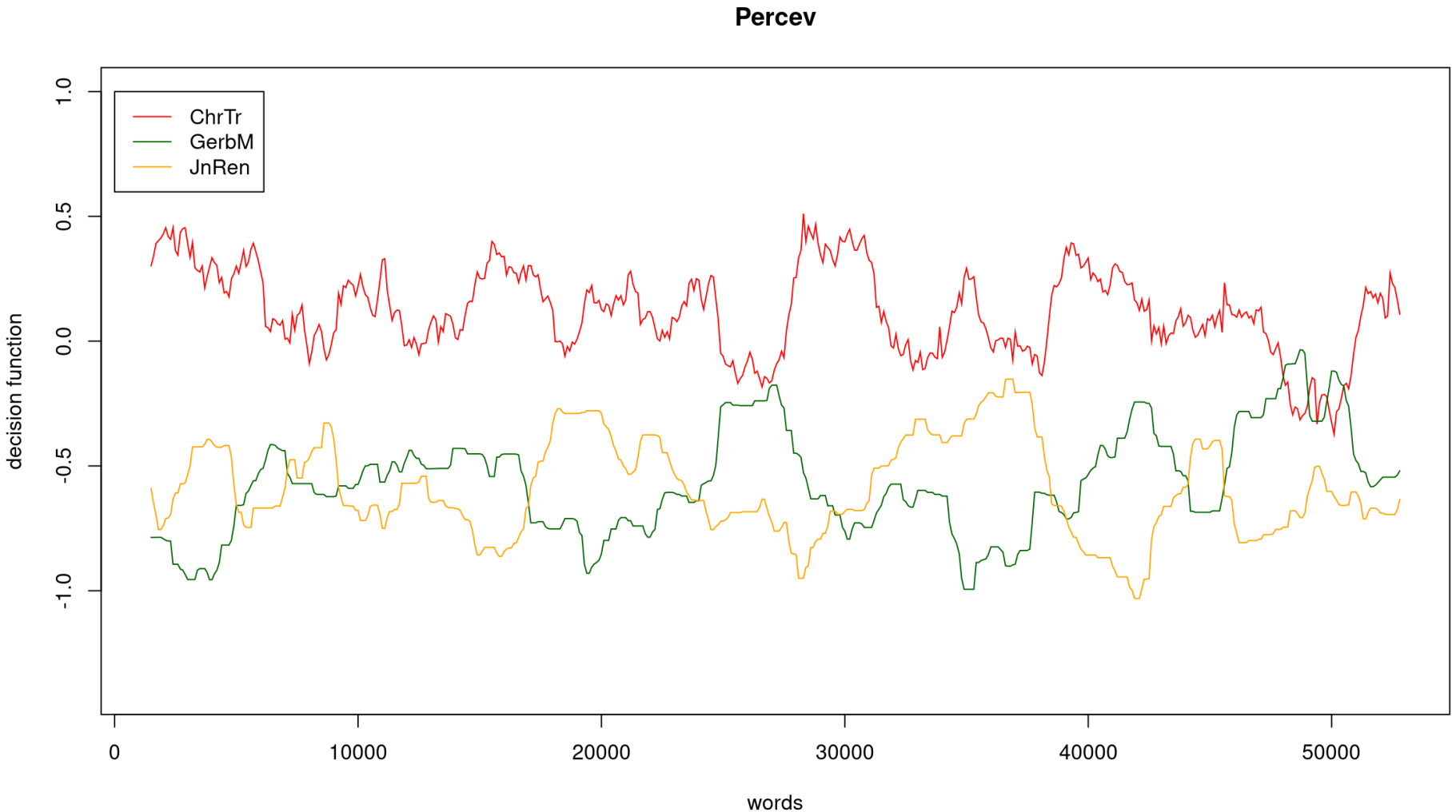


Outline

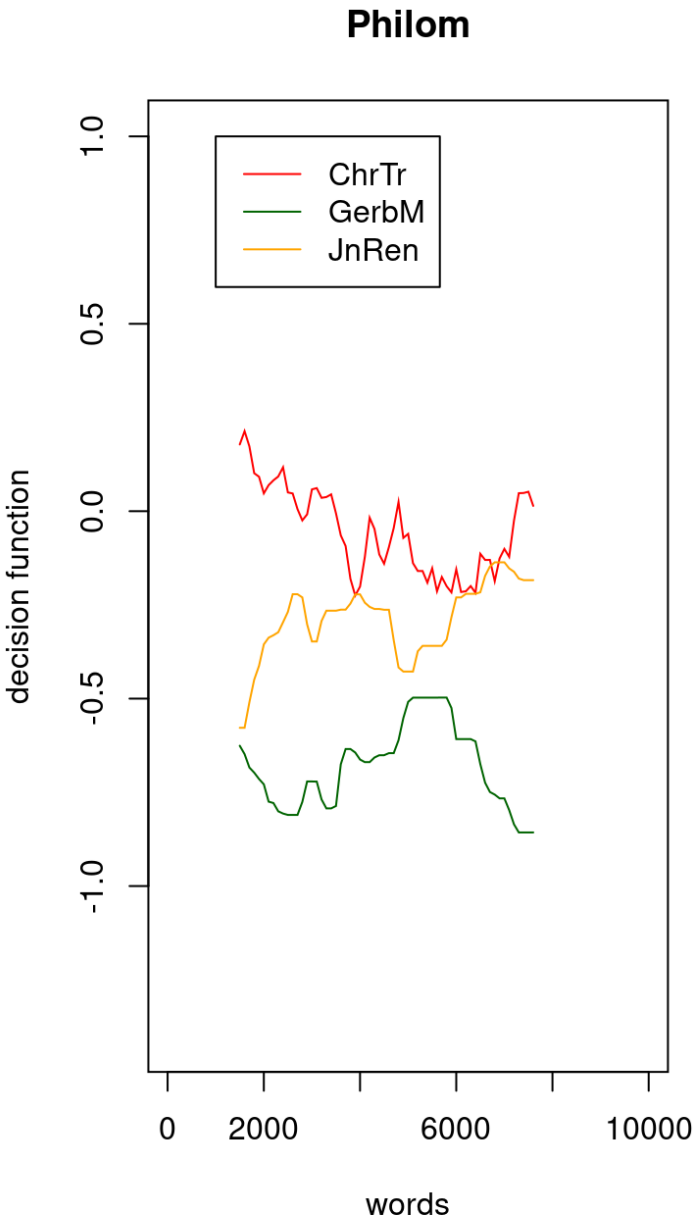
2024 Results: Lancelot



2024 Results: Perceval



2024 Results: Philomena



2024 Results: Guillaume d'Angleterre