Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1293–1302
Marseille, 11–16 May 2020

Automatic Orality Identification in Historical Texts

Katrin Ortmann, Stefanie Dipper
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
{ortmann, dipper}@linguistics.rub.de

Abstract

Independently of the medial representation (written/spoken), language can exhibit characteristics of conceptual orality or literacy, which mainly manifest themselves on the lexical or syntactic level. In this paper we aim at automatically identifying conceptually-oral historical texts, with the ultimate goal of gaining knowledge about spoken data of historical time stages. We apply a set of general linguistic features that have been proven to be effective for the classification of modern language data to historical German texts from various registers. Many of the features turn out to be equally useful in determining the conceptuality of historical data as they are for modern data, especially the frequency of different types of pronouns and the ratio of verbs to nouns. Other features like sentence length, particles or interjections point to peculiarities of the historical data and reveal problems with the adoption of a feature set that was developed on modern language data.

Keywords: conceptual orality, historical data, German

1. Introduction

Human language is used for communication in two major forms, written and spoken. Depending on the medium, utterances can differ significantly, as both discourse modes place different demands on the language user. While spoken discourse requires online processing, thus depending on working memory, written discourse may be processed multiple times and at any desired speed.

However, linguists recognize that besides this medial distinction there is also a lot of variation within discourse modes. As Halliday (1989, 32) explains, "'written' and 'spoken' do not form a simple dichotomy; there are all sorts of writing and all sorts of speech, many of which display features characteristic of the other medium."

This variation within discourse modes was termed *conceptual orality or literacy* by Koch and Oesterreicher (1985) and it can come in handy, if the historical development of certain linguistic phenomena related to discourse mode should be investigated. For historical time periods, obviously, only written language data is available. Hence, in order to gain knowledge about historical spoken discourse, it is necessary to identify written texts that are close to the oral mode, i.e. conceptually oral or spoken-like texts.

In this work, we test on historical German texts a set of general linguistic features that we have proven to be effective for the automatic identification of conceptual orality in modern language data (Ortmann and Dipper, 2019).

The remainder of this paper is structured as follows: Section 2. gives a short overview of the related work. Section 3. describes the historical data from different registers and Section 4. introduces the linguistic features used in this study. In Section 5. we use the features to classify the historical texts according to conceptual orality and inspect the results obtained both between and within registers. The paper concludes with a discussion of the findings in Section 6. and a summary in Section 7..

2. Related Work

The distinction between oral and literate language as introduced by Koch and Oesterreicher (1985) seems to be a general linguistic phenomenon and has been shown to play a role in typologically very different languages (Biber, 1995). The characteristics that Koch and Oesterreicher (1985) suggest to distinguish between oral and literate texts are too abstract and vague to be operationalized, though. Ágel and Hennig (2006), therefore, extend the approach by Koch and Oesterreicher (1985) and show how conceptual orality of modern as well as historical texts can be objectively assessed and measured using a range of linguistic features. However, Ágel and Hennig (2006)'s method is based on an in-depth manual inspection of every individual sentence to identify the linguistic features in a text. Thus, this method is not sensibly applicable to larger amounts of data.

To date, only few attempts were made to automatically identify conceptually oral texts. Rehm (2002) focuses on features that are specific to the domain of computer-mediated communication (CMC). In Ortmann and Dipper (2019), we propose a set of rather general linguistic features which we show to be effective for modern German texts from various registers. To test whether those features can also be used effectively for the identification of orality in historical texts, we will transfer the approach described in Ortmann and Dipper (2019) to historical German data.

3. The Data

For the present study we use historical German texts from Deutsches Textarchiv[1] (DTA, BBAW (2019)) from four different registers that exhibit different degrees of conceptual orality: specialist texts (*Science*), newspaper texts (*News*), narrative texts (*Fiction*), and funeral sermons (*Sermon*).[2] The

[1] Version from 2019, February 6, downloaded at `http://media.dwds.de/dta/download/dta-lingattr-tei_2019-02-06.zip`.

[2] Science: DTAmain class 'Fachtext', from a broad range of disciplines, e.g. biology, mathematics, medicine. News: DWDS class

Register	Example Sentence
Science	Zu dem Ende ist in 45) a = 1, b = 0 zu nehmen, hernach a und b für c und d zu schreiben. *To this end, in 45) a is to be taken as 1 and b as 0, thereafter a and b are to be written for c and d.*
News	Im ersten Artikel desselben wird bestimmt, daß die Budgets von 1839 während der ersten acht Monate des Jahrs 1840 in Kraft bleiben; nur bleiben die aus den ostindischen Geldmitteln genommenen 1,200,000 fl. von dem Einnahmebudget weg. *In the first article of the latter, it is stated that the budgets of 1839 remain in force during the first eight months of 1840; only the 1,200,000 fl. taken from the East Indian funds are excluded from the revenue budget.*
Fiction	O mein lieber Chamiſſo , ſelbſt vor Dir es zu geſtehen, macht mich erröthen . *O my dear Chamisso, even to confess it to you makes me blush.*
Sermon	Kein gröſſer ſchmertz auff Erden iſt/ Denn wenn der Tod mit gewalt aufflöſt Zwey Hertzen/ die in Lieb vnd Leid Feſt verbunden geweſen allezeit. *There is no greater pain on earth than when death by force separates two hearts that have always been united in love and suffering.*

Table 1: Example sentences from the four registers.

DTA is pre-annotated with automatically-created sentence and token boundaries, lemmas, and POS tags. Table 1 shows an example sentence from each register.

While funeral sermons are only available for the time range from 1550 to 1750, the vast majority of available fiction, newspaper and specialist texts was published between 1700 and 1900. We divided these 200-year spans into four time-windows of 50 years and randomly selected sentences summing up to approximately 100,000 tokens for each 50-year window, resulting in a total of 400,000 tokens for each register. Table 2 gives an overview of the data.

Expected orality Based on general characteristics of the registers, we try to assess the prototypical conceptuality of the texts to locate them on the literate-to-oral scale. Specialist texts normally deal with very specific topics and are written for meticulous reading by an intellectual audience, so they could be expected to be the conceptually most literate register in the study. Newspaper texts usually cover a broad range of topics, primarily trying to inform a large and diverse audience, so we expect them to be conceptually less literate than specialist texts. Fiction texts are written to entertain the reader and often contain a mix of dialogues and narrative passages, thus being more oriented towards conceptual orality. Likewise, we expect the funeral sermons to be conceptually oral as they were meant for oral presentation in front of a rather small audience and in a rather familiar atmosphere, although they might subsequently have been adapted or elaborated for printing.

It is important to notice that these assumptions are based on modern conventions. Degaetano-Ortlieb et al. (2019) showed that the conceptuality of a register can change over time: English scientific texts from past centuries used to be more orally-oriented than they are nowadays. So we might observe similar developments for the German data.

'Zeitung'. Fiction: DTAmain class 'Belletristik' without poetry and drama. Sermon: DTAsub class 'Leichenpredigt'.

[3]The DTA does not provide information on authorship with news texts.

Register	Years	#Tok	#Sent	#Doc	#Authors
Science	1700–1900	400,078	15,202	623	344
News	1700–1900	400,061	17,428	1070	– [3]
Fiction	1700–1900	400,042	15,841	309	128
Sermon	1550–1750	400,162	16,367	148	126
Total		1,600,343	64,838	2,150	

Table 2: Overview of the data used in the study, ranked by the expected orality of the registers from conceptually literate (Science) to conceptually oral (Sermon).

4. Features of Orality

In Ortmann and Dipper (2019), we used a range of general linguistic features which have been proposed in the literature as useful indicators of orality, such as the mean sentence length or the ratio of certain pronouns to all words. In the present study, we use almost the same features for the identification of orality, see Table 3 for an overview.[4]

As the historical texts are not annotated with syntactic dependencies, we exclude features that require dependency annotations, namely noun phrase complexity and the proportion of pronominal subjects. This leaves us with a total of 15 distinct features that relate to (syntactic) complexity, reference, lexicon and sentence type.

5. Results

As already mentioned, the features introduced in the previous section have been shown to be useful for the identification of conceptual orality and literacy in modern German texts (Ortmann and Dipper, 2019). To test whether they are equally useful for the historical data, we apply them to the four registers described above (Section 5.1.), before we look at developments within the registers (Section 5.2.).

[4]For a detailed description of the features, see Ortmann and Dipper (2019).

Feature	Description
Complexity	
mean_sent	Mean sentence length.
med_sent	Median sentence length.
mean_word	Mean word length.
med_word	Median word length.
subord	Ratio of subordinating conjunctions (tagged as KOUS or KOUI) to full verbs.
coordInit	Proportion of sentences beginning with a coordinating conjunction.
V.N	Ratio of full verbs to nouns.
lexDens	Ratio of lexical items (tagged as ADJ.*, ADV, N.*, VV.*) to all words.
Reference	
PRON1st	Ratio of 1st person pronouns with lemmas *ich* 'I' and *wir* 'we' to all words.
DEM	Ratio of demonstrative pronouns (tagged as PDS) to all words.
DEMshort	Proportion of demonstrative pronouns (tagged as PDS) with lemmas *diese* or *die* 'this/these' which are realized as the short form (lemma *die*).
Lexicon	
PTC	Proportion of answer particles (*ja* 'yes', *gewiss* 'certainly', *nein* 'no', *bitte* 'please', *danke* 'thanks') to all words.
INTERJ	Proportion of primary, i.e. one-word interjections (e.g. *ach*, *oh, o*, *bravo*, *halleluja*, *hmm*) to all words.
Sentence type	
question	Proportion of interrogative sentences, based on the last punctuation mark of the sentence.
exclam	Proportion of exclamative sentences, based on the last punctuation mark of the sentence.

Table 3: Features used for classification. The POS tags are from the STTS tagset. Punctuation marks are ignored except for the sentence-type features.

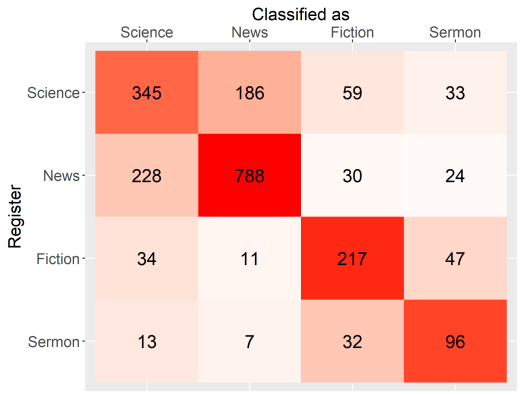

Figure 1: Confusion matrix for the classification of registers, summed over all cross-validations. Color intensity indicates the proportion of a register's classified texts per category.

Class	Precision	Recall	F-Score
Science	0.556	0.554	0.555
News	0.794	0.736	0.764
Fiction	0.642	0.702	0.671
Sermon	0.480	0.649	0.552
Weighted Avg.	0.682	0.673	0.676

Table 4: Results of classifying registers with the J48 decision-tree classifier.

5.1. Classifying registers

In Fig. 2, we plot the densities of the selected features in the four registers.[5] The plots show that many features clearly reflect the expected differences between the registers. For example, the mean word length is greater in conceptually literate registers than in rather orally-oriented registers. Also, the ratio of verbs to nouns is lower in Science and News than in the other registers. Science and News are the most literate registers, and they show a nominal style, in contrast to the more verbal style in the other registers. Further clear differences can be observed for the personal and (short) demonstrative pronouns, which are more frequent in the conceptually oral registers while they are mostly absent in the more literate registers. The same holds for particles, interjections, questions, exclamations and sentence-initial coordinating conjunctions.

Like in Ortmann and Dipper (2019), we train a J48 decision-tree classifier (Quinlan, 1993), which allows us to inspect the features used by the classifier to determine the registers.[6] A 10-fold cross-validation results in an overall accuracy of 67.26%, which is about 20 percentage points lower than for the modern data in Ortmann and Dipper (2019). Table 4 shows that the highest accuracy is achieved for the News register followed by Fiction, while classifying Sermon results in low precision and classifying Science in both low precision and low recall.

The confusion matrix in Fig. 1 indicates that confusions mostly happen among immediately neighboring registers, e.g. Science and News, which show similar levels of conceptual orality/literacy.

As observed in the plots, the pronoun features, the verb-to-noun ratio and the proportion of interrogative sentences are indeed very useful for the classifier in distinguishing the registers, which is evidenced by their information gain[7], cf. Table 5. It is interesting to note that sentence length is the least

[5]The plots have been created with the R package ggplot2, `https://github.com/tidyverse/ggplot2`.

[6]We use J48 as implemented in Weka (Witten et al., 2011) with the minimum number of instances per leaf set to 5, combined with a filter that balances the size of the different classes in the training data. So the options are set as follows:
`weka.classifiers.meta.FilteredClassifier -F "weka.filters.supervised.instance.ClassBalancer -num-intervals 10" -S 1 -W weka.classifiers.trees.J48 -- -C 0.25 -M 5.`

[7]Information gain is provided by Weka's "InfoGainAttributeEval" and is calculated as InfoGain(Class,Attribute) $= H(\text{Class}) - H(\text{Class}|\text{Attribute})$.

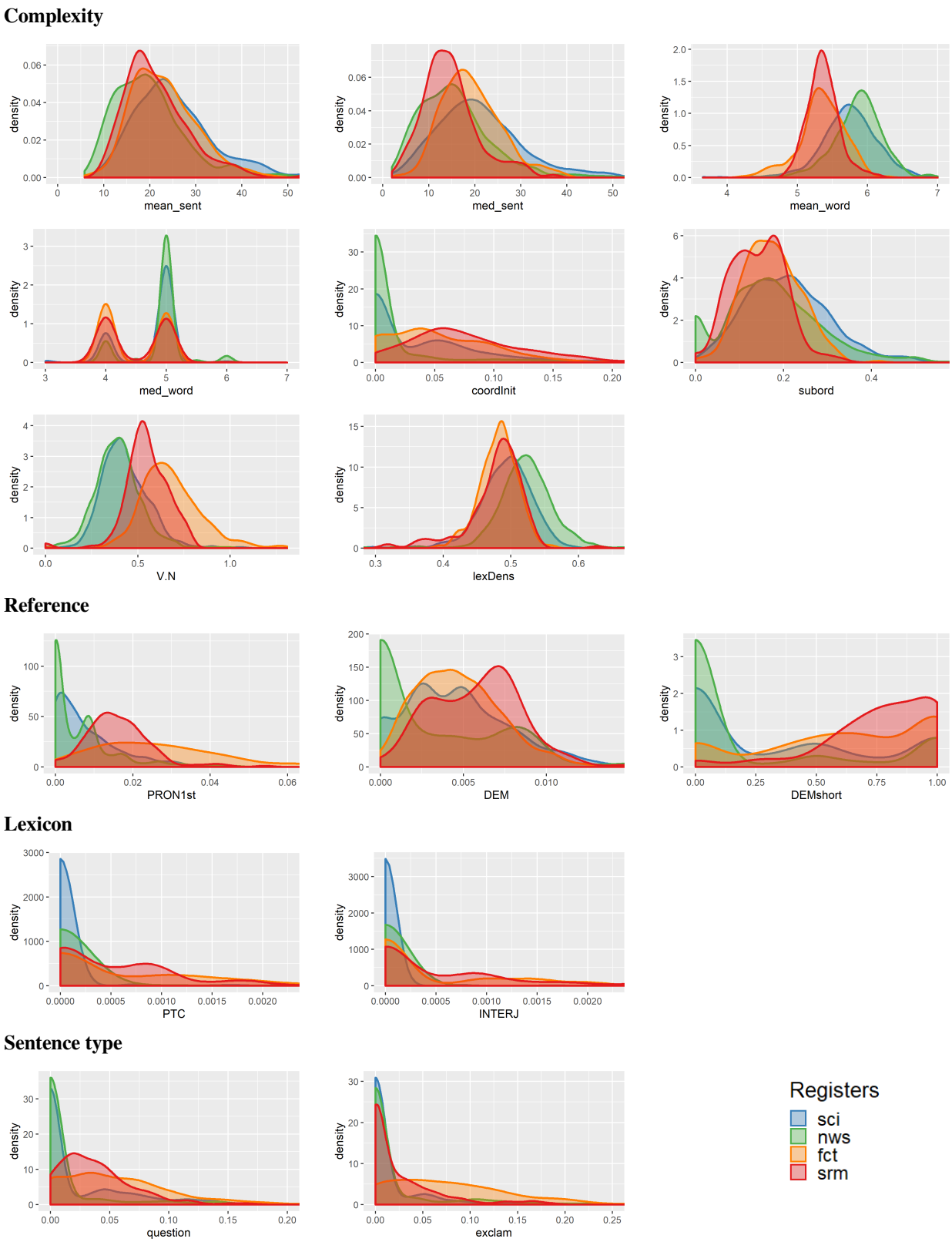

Figure 2: The graphs plot the densities of the features used in classification for each register. Registers are mapped onto color (blue: Science (*sci*), green: News (*nws*), orange: Fiction (*fct*), red: Sermon (*srm*)). Except for the features mean_word, med_word, V.N, and DEMshort, we set sensible cut-off points for the x-axis in each plot to exclude outliers.

Information Gain	Feature
0.361	DEM
0.322	V.N
0.313	PRON1st
0.256	question
0.237	DEMshort
0.234	mean_word
0.233	coordInit
0.235	PTC
0.193	exclam
0.193	lexDens
0.191	INTERJ
0.112	subord
0.103	med_word
0.093	med_sent
0.078	mean_sent

Table 5: Ranking of features according to their Information Gain with respect to registers.

relevant feature for the historical texts – contrary to the findings for modern data in Ortmann and Dipper (2019).

5.2. Historical Development within Registers

Each register covers a time span of 200 years (1550–1750 for Sermon, and 1700–1900 for the other registers), see Section 3.. As is already known from previous research, registers develop over time. For instance, for English scientific texts Degaetano-Ortlieb et al. (2019) observed a shift from oral orientation towards a literate style. Based on the same features as above, we test whether we can observe a development over time within the four registers. For this, we divide the 200-year spans into four 50-year time windows, each comprising 100,000 tokens.

An inspection of the feature distribution (cf. the density plots in Fig. 3–6 in the appendix) shows that texts in the Science register over time tend toward having longer sentences and a more nominal style, while, on the other hand, they contain less pronouns and (short) demonstratives. Perhaps this could be interpreted as a development to a more literate style, similar to the one observed by Degaetano-Ortlieb et al. (2019) for English scientific texts. For the News register, we can also observe a more nominal style and less (short) demonstratives over time. For Fiction the opposite is found: sentences and words become shorter, there is less subordination, but more questions and exclamations. Likewise there are more (short) demonstratives, interjections and particles, which could indicate a more oral-oriented conception of this register. For Sermon no clear development can be recognized. Overall, sentences in the latter register seem to become shorter with more subordination and less sentence-initial coordination. Also there are more questions and exclamations, but a more nominal style with less pronouns.

Training a J48 decision-tree classifier to predict the four 50-year time windows in each register results in low F-scores $<$ 53.2% (weighted average over the 50-year windows).[8] This shows that the registers are rather homogeneous and any development within the registers is too subtle to be reliably detected by the features considered in the classification.

6. Discussion

As the results from the previous section show, most of the features are equally useful for the classification of historical texts as for modern data, especially features of reference and deixis (realized as different types of pronouns), the ratio of verbs to nouns and the sentence type. In contrast to the findings in (Ortmann and Dipper, 2019), the simple feature of sentence length, which proved to be useful for modern data, is the least helpful feature in the classification of our historical data.

However, the classification accuracy for historical data (67.26%) is much worse than for modern data (88.28%). One possible reason could be that the set of features, although already very general, is still too much tailored towards modern conventions and needs to be adjusted to better fit the historical data. For example, some features like interjections and particles are typical characteristics of modern conceptually-oral texts but are extremely rare in all of the selected historical registers. On the other hand, there might be features which are highly relevant for historical data but less so for modern data, some of which may require further annotations that are not available in the DTA, e.g. syntactic dependencies.

Besides the feature set, a manual analysis of the data suggests that the results could also be negatively influenced by the annotation quality. For instance, the POS annotation of some pronoun types in the DTA seems to be problematic, as for example demonstratives are often confused with articles and relative pronouns. Also, the automatically created sentence boundary annotations in the DTA seem to be incorrect in many cases, resulting in very long or very short false sentences. In addition to that, the general concept of sentencehood and the use of punctuation in general has changed over time (from the use of the virgule '/', marking pauses or arbitrary syntactic units, to modern punctuation conventions), which has an impact on features of sentence length and sentence type. So, more accurate annotations might improve the results of the automatic identification of orality.

Finally, another possible explanation for the lower accuracies could lie in the data itself. Some text samples contain fragments or even entire sentences in Latin or French and, hence, cannot be properly analyzed unless the foreign language material is either included in the analysis as an additional feature or filtered out before the analysis. Moreover, it could also be the case that the differences between our historical registers are just not as extreme as between the modern registers compared in Ortmann and Dipper (2019) (Chat, Dialog, TED, Speech, News), making the historical registers harder to classify. Or it could be a consequence of data sparseness, in that

[8] An exception is the News register, with an F-score of 75.6%. However, the high value stems almost only from the high results for a single time window (1800–1850: weighted average F-score 89.2%), which comprises two thirds of the texts in this register (with the same total amount of tokens, though) and is thus over-represented in the weighted average. If the number of texts is not taken into account, the F-score for the classification within the News register lies below 60%.

very heterogeneous registers (e.g. specialist texts stemming from a wide range of scientific disciplines), and the large time windows of 200 years result in too much variation *within* the registers, obfuscating the differences *between* registers. In this case, a next step could be to either narrow down the registers or to only predict single historical 50-year windows and test whether this works equally well as for modern data using only 100,000 tokens per register.

7. Summary

For historical time periods only written language resources are available. Therefore, for the investigation of phenomena related to discourse mode and their historical development, conceptually oral, i.e. spoken-like, texts need to be identified. In this paper, we tested a set of general linguistic features we proposed in Ortmann and Dipper (2019) to automatically identify conceptual orality in historical texts. Our analyses show that many of the features are indeed equally useful in determining the conceptuality of historical data as they are for modern data, especially the frequency of different types of pronouns and the ratio of verbs to nouns.
However, some features like sentence length, particles or interjections also point to problems with the adoption of a feature set which was developed on modern data. As a consequence, we discussed how peculiarities of historical data and the available annotations might influence classification results and which steps could be taken to improve the automatic identification of conceptual orality in historical texts.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 (Project C6).

8. Bibliographical References

Ágel, V. and Hennig, M. (2006). *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000*. Niemeyer, Tübingen.

BBAW. (2019). Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften; `http://www.deutschestextarchiv.de/`.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., and Teich, E. (2019). An information-theoretic approach to modeling diachronic change in scientific English. In Carla Suhr, et al., editors, *From Data to Evidence in English Language Research*. Brill, Leiden, NL/Boston, MA.

Halliday, M. A. K. (1989). *Spoken and written language*. Oxford University Press.

Koch, P. and Oesterreicher, W. (1985). Sprache der Nähe — Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Ortmann, K. and Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. In *Proceedings of the NAACL-Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 64–79, Minneapolis, MN, USA.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Rehm, G. (2002). Schriftliche Mündlichkeit in der Sprache des World Wide Web. In Arndt Ziegler et al., editors, *Kommunikationsform E-Mail*, pages 263–308. Stauffenburg, Tübingen. Retrieved from `http://www.georg-re.hm/pdf/Rehm-Muendlichkeit.pdf`.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA, 3rd edition.

9. Appendix

Fig. 3–6 show density plots comparing the distribution of each of the 15 features used in the study across four time windows (with registers Science, News, and Fiction: 1700–1749, 1750–1799, 1800–1849, 1850–1899; with register Sermon: 1550–1599, 1600–1649, 1650–1699, 1700–1749). For some features, we set sensible cut-off points for the x-axis to exclude outliers (see figure captions).

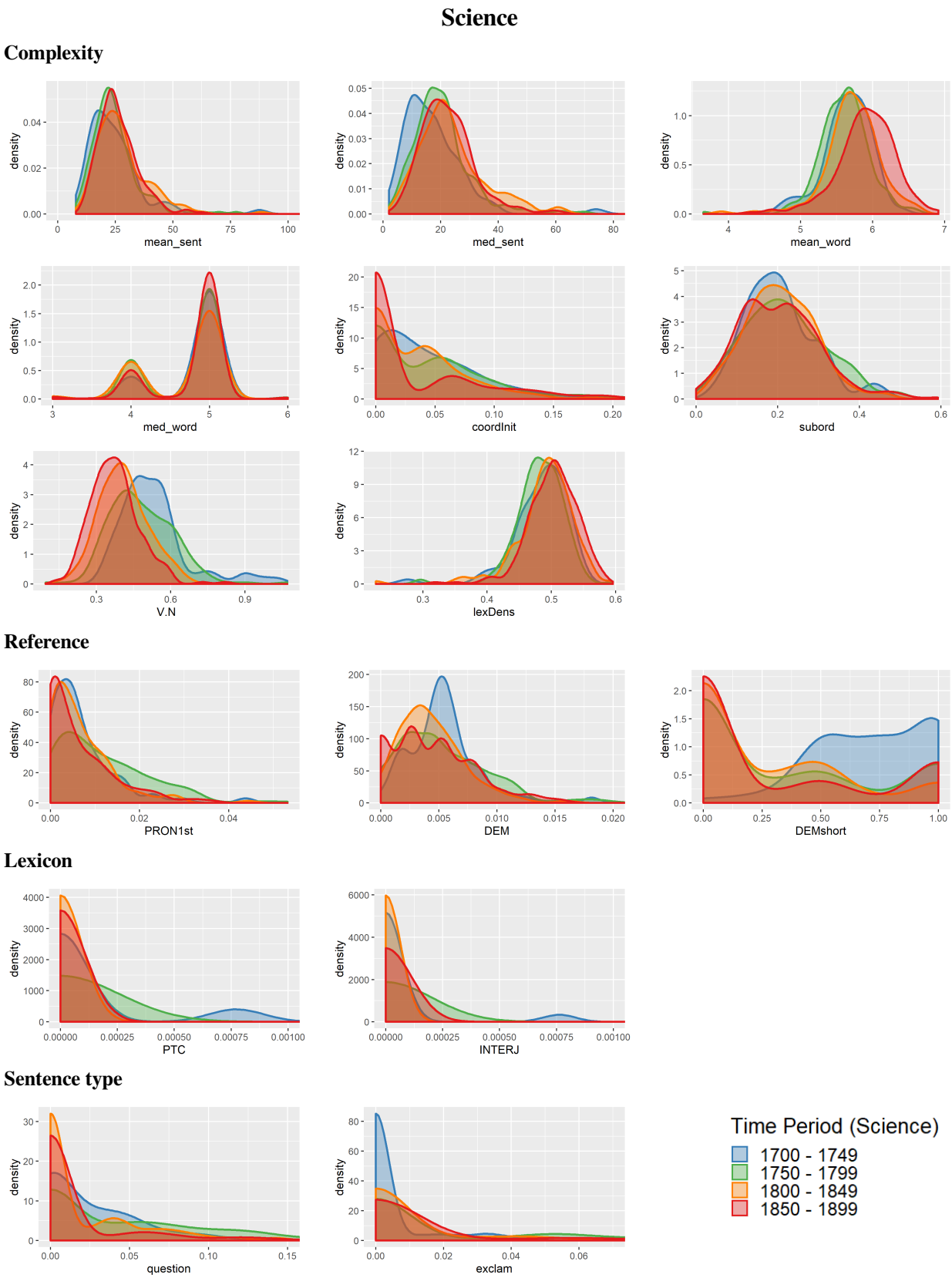

Figure 3: Density plots for the **Science** register within the four 50-year time windows. Cut-off points were set for the features mean_sent, med_sent, coordInit, DEM, PTC, INTERJ, question and exclam.

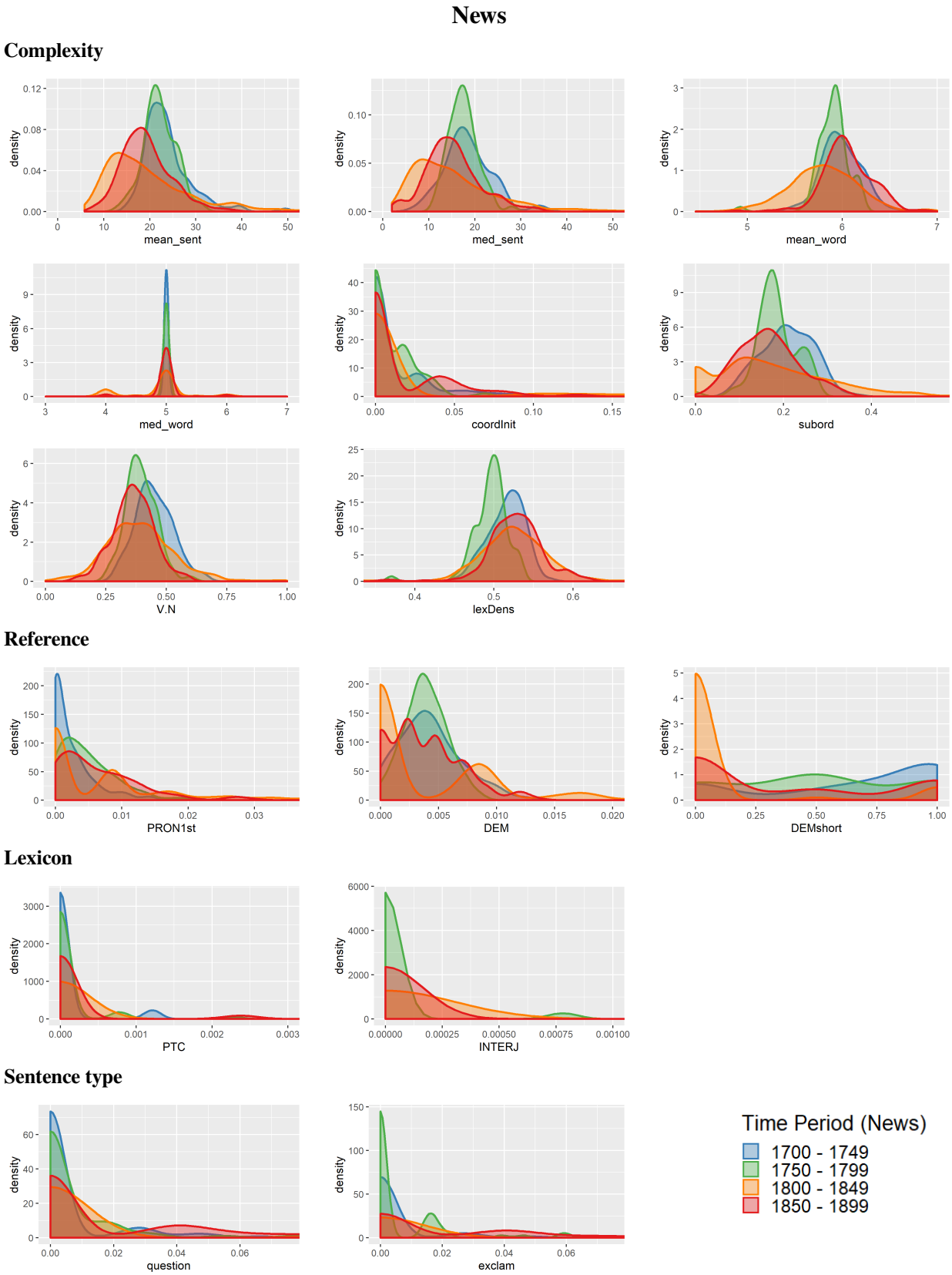

Figure 4: Density plots for the **News** register within the four 50-year time windows. Cut-off points were set for all features except mean_word, med_word and DEMshort.

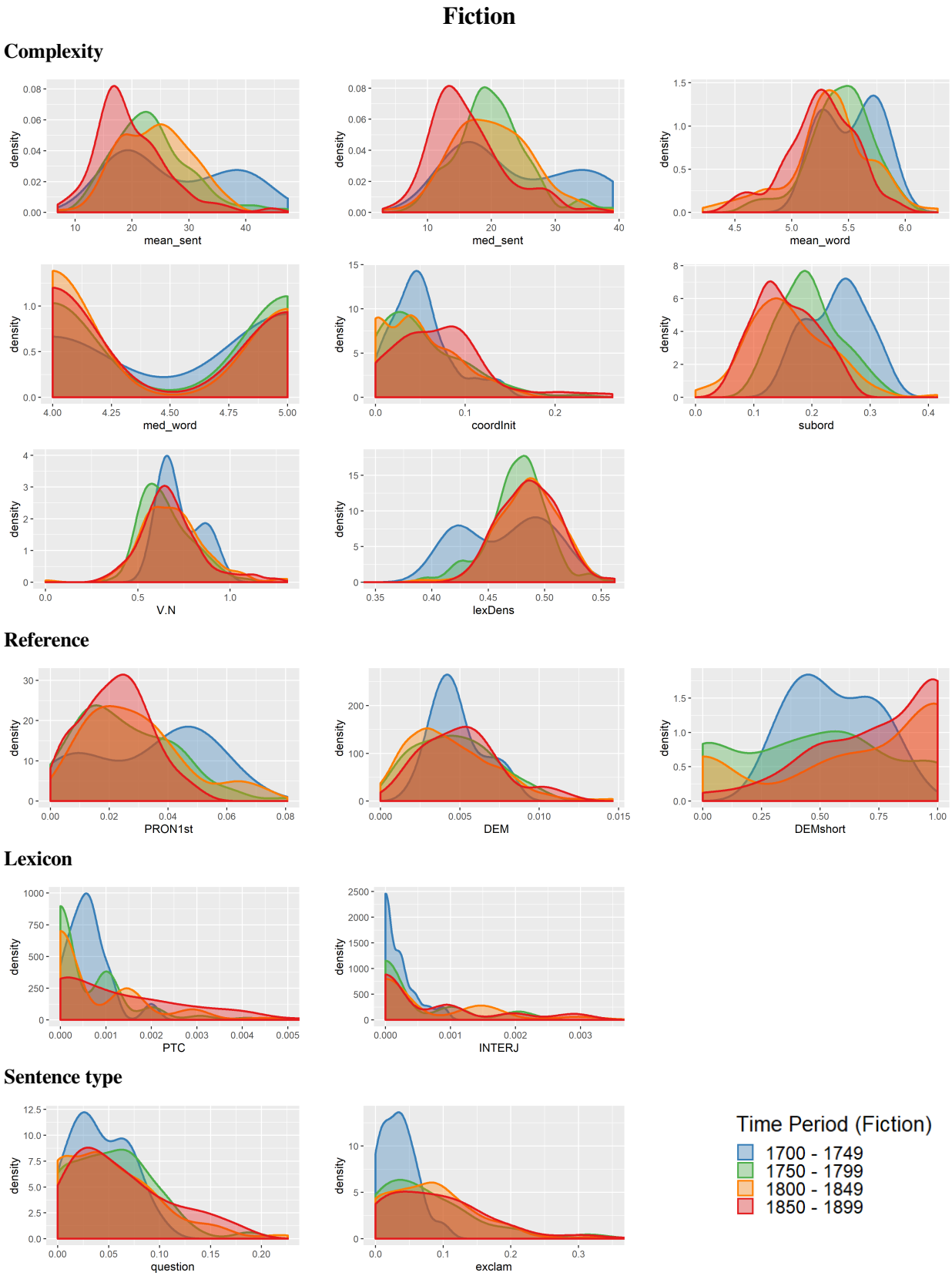

Figure 5: Density plots for the **Fiction** register within the four 50-year time windows. Cut-off points were set for the features lexDens, PTC, INTERJ and exclam.

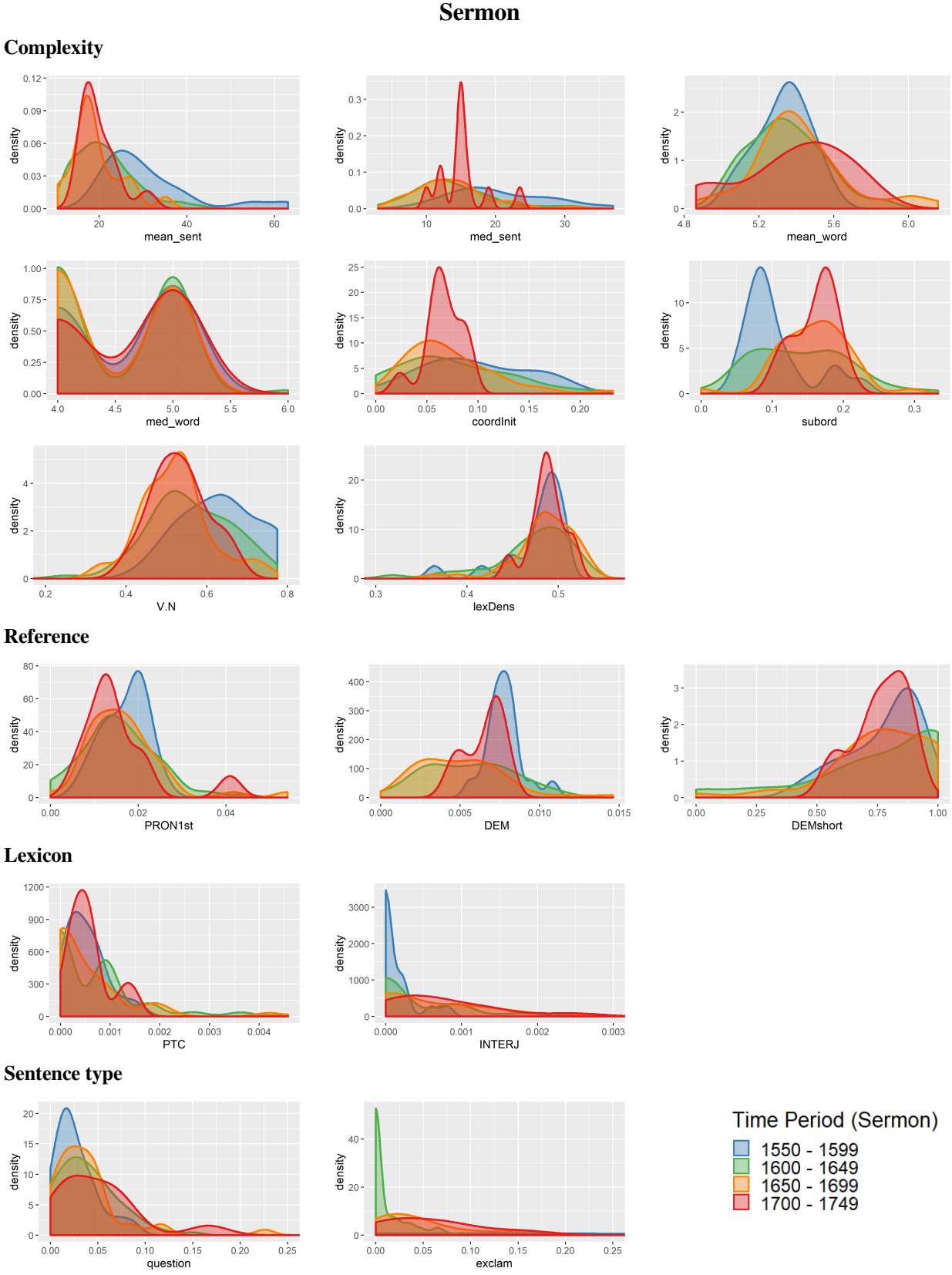

Figure 6: Density plots for the **Sermon** register within the four 50-year time windows. Cut-off points were set for the features V.N, lexDens, INTERJ, question and exclam.