A Dataset of Historical Medical Periodicals Annotated with Textual Genre

Vera Danilova[1] **and Sara Stymne**[2]
[1]Department of History of Science and Ideas, Uppsala University, Sweden
[2]Department of Linguistics and Philology, Uppsala University, Sweden
vera.danilova@idehist.uu.se, sara.stymne@lingfil.uu.se

Abstract

Historical corpora, especially those compiled from magazines and periodicals, are complex due to the diversity of text types and evolving genre conventions. Addressing these challenges requires systematic genre annotation and well-defined classification schemes to support downstream NLP tasks. This paper introduces a dataset of historical medical periodical texts in German and Swedish annotated for textual genre and additional features that may influence genre identification, such as the presence of OCR errors. We describe the development of the genre classification, annotator recruitment and training procedures, and provide an analysis of the annotator agreement.

Keywords: Genre Annotation, Historical NLP, Historical Medical Periodicals, Digital Humanities

1. Introduction

Historical periodicals are highly heterogeneous materials that reflect the evolution of both printing technologies and communicative practices. They encompass multiple text types and genres, which changed over time alongside increasingly diverse and complex layouts. A substantial body of such historical magazines has already been digitized by the digital humanities community across various domains. However, the temporal, linguistic, and task-related coverage of these resources still varies considerably. Consequently, the development of new evaluation datasets to support downstream NLP tasks and enhance the overall quality of research in historical NLP remains an important goal.

The ActDisease project[1] contributes to this effort by digitizing and developing tools for historical medical periodicals. The project focuses on the history of twentieth-century European patient organizations and employs mixed methods to analyze a wide range of materials, including patient organization magazines, annual reports, and medical journals.

In this context, genre segmentation serves not only as a means of dividing the material into coherent textual units but also as a way to study the evolution of communicative strategies within these sources. However, existing genre classification schemes vary widely, are often designed for different domains, and fail to capture genres specific to medical periodicals.

This paper presents a new dataset of OCR-processed texts in German and Swedish, annotated with textual genres and supplementary attributes, such as the presence of OCR errors, which may influence genre classification. In traditional annotation projects, the goal is to minimize disagreement between annotators, typically followed by reaching a single gold-standard by expert adjudication or aggregation (see e.g. Ide and Pustejovsky (2017) for an overview). However, recently, there has been an increased awareness of that variation in annotations can be due to valid underlying ambiguity rather than errors or fuzzy guidelines (Plank et al., 2014; Basile et al., 2021). In light of this, our dataset contains the annotations from each annotator, and we provide an analysis of annotator differences, related to genre labels.

In this paper, we provide a genre inventory targeted at historical medical periodicals, and use it to annotate data from four periodicals. We describe the annotation procedure in detail and provide an in-depth analysis of inter-annotator agreement. We propose a framework for analyzing disagreements across genre labels with respect to syntactic, semantic and topic-based features, that we use to further explore annotation differences. The code and annotation guidelines are available on GitHub[2].

2. Existing Genre Datasets and Annotation Schemes

There are a very limited number of historical collections annotated with textual genre. A recent work by Stahel et al. (2025) describes the annotation of a dataset of 7k English-language articles from 106 historical newspapers (1839-1903) with soft genre labels on article level. The genre scheme is custom and based mainly on the prior knowledge of the researchers about the dataset, as well as additional information from previous research on newspaper genres, the inventory of article categories and other

[1]ERC-2021-STG-101040999, Dept. of History of Science and Ideas, Uppsala University, Sweden

[2]https://github.com/ActDisease/genre-annotation-project

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 973–984
11–16 May 2026.

information. The authors report the Inter-Annotator Agreement (IAA) Krippendorff's $\alpha = 0.66$ computed based on the primary (best-fit) genre labels.

Harbers (2014) describes the manual coding of Dutch, English, and French historical newspapers into 18 journalistic genres, resulting in a metadata collection (no digitized editions were available). Genre categories were based on historical journalistic practice, handbook definitions, and newspapers' self-classifications, with labels assigned at the article level. The codebook, was applied by two groups of bilingual coders—one for Dutch and English, and one for French. While the authors note that both students and project researchers participated, the number of coders is not specified. IAA (Krippendorff's α) was 0.67 and 0.83 for the two groups, respectively. In (Bilgin et al., 2018), part of this metadata was later linked to a digitized Dutch periodical, yielding 1424 articles used as a gold standard for genre classification.

In the web genre domain, stable genre schemata have long been the focus, and several publicly available datasets have been produced. The largest and most comprehensive ones are CORE (Egbert et al., 2015) and FTD (Sharoff, 2018), which serve as the basis for neural genre classification models. In both, the annotation is on the document level. The IAA (Krippendorff's α) is 0.66 (main categories) and 0.76, respectively.

Unlike previous studies that annotated genres at the document level, our dataset applies genre labels at a more fine-grained level—the paragraph. The annotation schema accounts for genres specific to our historical medical periodicals, while also incorporating categories from earlier schemata developed for comparable materials.

3. Genre Categories

There is no universal genre scheme that can be readily applied across all contexts. Historical datasets illustrate this complexity particularly well, as they reflect how genres evolved over time and mirrored the communicative strategies of editors and authors.

In this project, our primary goal is to segment historical medical magazines where genres[3] are specific to this material type. These genres served as vehicles for different kinds of audience interaction — for example, promoting products (advertisement), informing about events (announcement), advocating certain viewpoints (argumentative), or providing entertainment (fiction, quizzes, humour).

In the context of our dataset of historical medical periodicals, genre classification became a practical means of segmenting materials, allowing texts with similar communicative functions to be analyzed together. The creation of our genre scheme was guided both by insights from historians of medicine in the project and by established genre and communicative-purpose classifications (Kuzman and Ljubešić, 2023; Caselli et al., 2022).

We distinguish the following genres in the dataset:

1. **Academic**: *Inform or report* on research in an accessible way, often referencing prior work.
2. **Administrative**: *Report* on administrative initiatives in healthcare, social services, or governance (outside of Patient Organizations).
3. **Advertisement**: *Promote* a product or service with intent to sell.
4. **Announcement**: *Inform or report* about the upcoming events and activities within the organization. Unlike invitations, announcements are objective and informative; they do not include promotional elements or encourage readers to make purchases.
5. **Appeal**: *Call to action* without a commercial goal, request for support, typically for a non-profit campaign (e.g., appeals for donations or participation).
6. **Argumentative**: *Express opinion or viewpoint* with intent to persuade the readers, typically authored by editors or contributors.
7. **Bio**: *Narrate or report* a life story or personal experience (often patient stories).
8. **Fiction**: *Entertain or engage emotionally* (short stories, poems, humour pieces).
9. **Guidance**: *Instruct or recommend*, e.g., dietary advice, recipes, manuals.
10. **Informational**: *Explain* a concept factually, similar to encyclopedia entries or definition.
11. **Interactive**: *Engage or entertain*, e.g., quizzes, crosswords, chess parties.
12. **Invitation**: *Invite or promote* participation in an event (e.g., patient cruises), contains promotional elements and descriptions with intent to sell.
13. **Legal**: *Explain* legal terms, regulations, or policies.
14. **News**: *Report* on a recent event (what happened, when, and where), outside of the patient organization domain.
15. **QA**: *Explain or resolve doubts* in question–answer sections, e.g., "Frågan är fri" ("The question is free").
16. **PO_report**: *Report or narrate* activities of patient organizations or publication updates.

4. Annotator Recruitment

We recruited eight annotators, four each for German and Swedish, for up to 75 hours of annotation in a month, asking them to annotate as much as

[3]We use the definition of genre as a group of texts that share a communicative purpose (Kessler et al., 1997).

they could within that time. The annotators were compensated according to university guidelines. A call for applications was circulated via mailing lists of Humanities departments, with the main requirement being proficiency in either German or Swedish.

Pairs of annotators were given the same material. One Swedish annotator later withdrew and could not be replaced, resulting in a portion of the Swedish dataset being annotated by a single annotator. This single annotator has a strong background in benchmark annotation for AI training. The overview of annotators' background is provided in Tab. 1. Annotators were paid by the hour up to a maximum number of hours. Because annotators had different speeds and many did not reach the maximum number of hours allocated, the number of annotations per annotator varied.

We conducted separate training sessions for the Swedish and German annotators to introduce the annotation procedure. During these meetings, we reviewed the guidelines, hosted on a shared Notion workspace, and clarified their application by discussing challenging examples. The workspace was continuously updated with answers to the annotators' questions and additional information. Annotators were encouraged to communicate with project researchers throughout the process.

5. Annotation Procedure

5.1. Sampling

To ensure temporal and topical diversity, we sampled pages from various time periods of several Swedish and German medical journals: 1) Diabetiker Journal (1951–1990), published by the German Diabetes Association; 2) Der Allergiker (1959–1985), published by the German Association for Allergy and Asthma; 3) Diabetes (1949, 1952–1990), published by the Swedish Diabetes Association; 4) Status (1938–1991), published by the Swedish Association for Lung Diseases.

Materials from Diabetes were included for both Swedish and German because the collaborating historians of medicine were already familiar with these publications and their text types. Their expertise can further provide valuable insight into interpreting and validating the genre classification results. The Status periodical was included due to its greater genre diversity – particularly a higher proportion of entertainment-related genres – thus contributing to a richer representation of text types within the dataset.

Overall, historians assessed the genre distributions to be broadly comparable across the periodicals. Nevertheless, individual publications exhibit systematic preferences for particular genres, a pattern that is also reflected in the sampled data. For instance, the German Diabetes Association periodical predominantly features academic texts, whereas its Swedish counterpart focuses largely on organisational and administrative reporting. Such naturally occurring imbalances make it difficult to obtain a uniform genre distribution across languages.

In contrast to title-based candidate selection methods (e.g., Stahel et al. (2025)), we adopted a diversification strategy based on temporal and issue-level variation, sampling pages from different time periods within each periodical. Concretely, for each journal, the material was divided into two time periods: pre-1960 and post-1960. Since 1960, layouts of periodicals have become significantly more complex, dynamic, and visually driven. For each periodical, 200 pages were randomly sampled from each time period. Annotation was carried out by two groups per language, with each group assigned to one of the two time periods (pre-1960 vs. post-1960).

Each group of annotators was provided with files containing text segments extracted from XML outputs generated by the ABBYY OCR engine.[4] These XML files encode detailed, multi-level recognition results, where elements ranging from characters to lines, paragraphs, and blocks (i.e., bounding boxes encompassing multiple paragraphs) are enriched with formatting attributes such as font size and font face.

Preliminary experiments with different segmentation levels revealed that block-level units were overly coarse, often merging paragraphs from distinct genres. In contrast, paragraph-level segmentation offered a more coherent and reliable unit of analysis. Accordingly, we extracted paragraphs with their associated formatting attributes (font size, font face, bold, italic). Paragraphs sharing identical formatting patterns—defined as a full match across these attributes—were subsequently merged.

Due to the complexity of page layouts, the OCR process was unable to consistently recover complete article text. The resulting output was noisy, both in terms of orthographic accuracy and reading order. Paragraph-level annotation, therefore, provided a practical compromise, corresponding to the smallest reliable unit obtainable from the OCR output. Moreover, performing genre classification at the paragraph level facilitates the reconstruction of semantically coherent article units, which can improve OCR post-processing and support downstream NLP applications.

Metadata were preserved, including file name, periodical title, year, volume, issue, and page num-

[4] ABBYY FineReader 14 Server: https://help.abbyy.com/en-us/finereaderserver/14/help/introduction/

No.	Course / Degree	Relevant Experience	Language
1	M.A. in Language Technology	Benchmark annotation for AI training	Swedish (native)
2	M.Sc. in Organic Chemistry	Medical periodicals	German (native)
3	M.A. in Language Technology	Proofreading and translation	German (C1)
4	B.A. in Linguistics	—	German (native)
5	M.A. in Digital Humanities	Annotation and periodicals	German (native)
6	M.A. in Linguistics	Translation	Swedish (native)
7	M.A. in History of Ideas	Transcription	Swedish (native)

Table 1: Background information of annotators.

ber. Paragraphs were presented in their original page order to maintain contextual coherence, and duplicate entries were removed.

5.2. Annotation Framework

Annotators accessed detailed guidelines via a shared document, which included:
(1) annotation fields and color coding;
(2) decision trees illustrated with diagrams;
(3) workflow instructions;
(4) detailed genre definitions; and
(5) examples.

Additional sections provided logistical information, answers to annotator queries, and sample page images.

We designed the annotation framework to be as simple and transparent as possible. Each annotator received an individual spreadsheet file in which semantic blocks were visually highlighted using color coding. To reduce cognitive load and maintain consistency, we employed hard assignments (binary 1/0 labels) rather than graded or probabilistic annotations. This choice was also motivated by the observations that, although soft labeling schemes are sometimes introduced in similar tasks, in practice, a single-consensus label is often produced, and no systematic benefit from soft labeling is observed (Jamison and Gurevych, 2015).

5.3. Procedure

Annotators were instructed to assign exactly one genre category (see Section 3) to each paragraph by marking 1 in the relevant genre column (and 0 otherwise). If most of a paragraph corresponded to a primary genre, that genre received the label 1. In cases where it was too difficult to identify a genre, annotators could leave the genre fields blank. To help annotators maintain contextual awareness, each file contained sequences of consecutive pages from a single issue, allowing them to refer back to the broader textual setting during annotation. Annotators were instructed to assess the degree to which a paragraph's genre was independent from that of the surrounding text. If a paragraph clearly stood out as belonging to a different genre – exhibiting its own distinct communicative function that differed from neighboring paragraphs – they were instructed to assign it that specific genre label.

They also annotated several auxiliary attributes:

- **ocr_errors** Contains OCR errors (misspellings).
- **contents_page** Is a table of contents.
- **publication_info** Metadata (editorial information).
- **art_author** Contains an article author's name.
- **art_title** Is an article title or subheading.
- **opinionated** Contains opinion (but not *Argumentative*).
- **dialogue** Multi-speaker discussion (e.g., interview).
- **caption** Is a caption (e.g., "Foto: Lala Aufsberg").

After the first 300 annotations, annotators provided feedback on the process. They reported that genre categories were generally clear and easy to apply, but that auxiliary attributes — especially *caption* — were sometimes difficult to determine due to limited context or OCR layout issues. In such cases, annotators were provided access to original page images for clarification.

Annotators were also encouraged to leave comments in a special column if uncertain about a paragraph's label, particularly when encountering hybrid genres. They were asked to indicate which genres co-occurred according to their judgment. Annotators processed their files individually and did not discuss the annotation process between themselves[5].

6. Description of the Resulting Dataset

The dataset contains a total of 23,031 unique labeled paragraphs (both single and double-annotated). In the first iteration of the annotation, there are 17,439 single-annotated paragraphs and 5592 double-annotated paragraphs. For each language, the data were annotated by two independent groups. Within each group, both annotators

[5]The shared document (Notion workspace linked from the GitHub repository) included example paragraphs to assist annotators in understanding the task.

Language	Group	Annotator (Size)	μ	σ
German	1	4 (1304), 3 (4443)	30.6	45.5
	2	2 (816), 5 (3405)	41.0	47.0
Swedish	1	1 (10362)	33.7	38.4
	2	7 (3788), 6 (4505)	23.8	29.7

Table 2: Number of labeled paragraphs (Size) per language, group and annotator. μ and σ correspond to the average paragraph length in tokens and its standard deviation, respectively.

worked on the same set of files to enable IAA analysis. The distribution of labeled data across languages, groups, and annotators is shown in Tab. 2. As the table indicates, the number of annotations varies considerably among annotators. Within each group, the annotations completely overlap: the annotator with the smaller set labeled a subset of the sentences annotated by the other annotator. Annotator No.1, who reported having the strongest professional background in annotation, contributed the largest portion of the labeled material. However, since their assigned annotation partner was unable to participate, these annotations are all single-annotated.

Tab. 2 also reports the distribution of paragraph lengths in tokens (based on whitespace) in each group for both languages (μ and σ). In German, the average paragraph length is slightly higher in group 2 than in group 1 (41 vs. 30.6). In Swedish, the paragraphs are on average shorter in group 2 compared to group 1 (23.8 vs. 33.7).

Tab. 3 shows the number of labeled paragraphs for each periodical (double and single-annotated). Annotator No.1 (without a pair) labeled a large number of Diabetes paragraphs (9376), which led to the highest coverage of the Diabetes periodical in the dataset. Moreover, Annotator No.1 annotated 986 Status paragraphs, while group 2 provided a balanced coverage of its Diabetes sample (4505 and 3788 paragraphs).

All annotators contributed to labeling paragraphs for the presence of OCR errors (binary label: error vs. no error). In total, 1127 paragraphs containing OCR errors were identified. In the German material, the number of detected errors is 1.7 times higher than in the Swedish material. Advertisements account for the largest share of OCR errors (30% of all detected errors), followed by Informational and Academic texts, which account for 21% and 10%, respectively.

6.1. Inter-Annotator Agreement

IAA was calculated separately for each group using standard coefficients (Artstein and Poesio, 2008). These metrics are appropriate for assessing agreement between two annotators assigning categorical labels: simple (observed) percentage agreement, the chance-corrected Cohen's κ[6], and Krippendorff's α[7].

Language	Periodical	Size (s / d)
German	Der Allergiker	2702 / 1127
	Diabetiker Journal	3122 / 945
Swedish	Diabetes	10629 / 3520
	Status	986 / —

Table 3: Total number of paragraphs by language and periodical (s – single / d – double annotations).

Language	Group	α	κ	%	Size
Swedish	2	0.67	0.67	73	3.5k
German	1	0.64	0.64	69	1.3k
German	2	0.46	0.47	57	778

Table 4: IAA per Language and Group: α - Krippendorff's Alpha, κ - Cohen's Kappa, and % - percentage agreement.

Tab. 4 shows agreement per annotator group. The Swedish group achieved α of 0.67 and κ of 0.67, corresponding to moderate and substantial agreement respectively, according to standard interpretations (Marzi et al., 2024; Artstein and Poesio, 2008). German Group 1 exhibited slightly lower levels of agreement, whereas Group 2, with a substantially smaller amount of material, had lower agreement numbers.

In our study, we acknowledge the inherent subjectivity of genre annotation, as noted in previous research (Kuzman and Ljubešić, 2023; Stahel et al., 2025), and the added difficulty posed by paragraph-level annotation of historical, domain-specific material. Annotators also reported that longer paragraphs were generally harder to label consistently. Previous work on genre annotation (Stahel et al., 2025; Harbers, 2014) have reported α values of 0.66–0.83 across several languages, being relatively comparable to ours. These findings suggest that agreement levels in genre annotation tasks can fluctuate considerably across languages and datasets, reflecting both linguistic variation and the differing interpretive challenges inherent to historical texts.

We hypothesized that IAA would vary substantially across genres, and to examine this, we computed Cohen's κ for each genre within each group. The results are presented in Tab. 5, which also includes the number of annotations per annotator and genre.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[7] https://pypi.org/project/krippendorff/

Genre	Lang/Group	Cohen's κ
Academic	DE1	0.27 (16/41)
	DE2	0.32 (33/115)
	SE2	**0.75** (211/188)
Administrative	DE1	0.27 (3/19)
	DE2	0.00 (6/0)
	SE2	0.36 (21/94)
Advertisement	DE1	**0.88** (303/313)
	DE2	**0.87** (130/137)
	SE2	**0.77** (685/485)
Announcement	DE1	0.38 (63/49)
	DE2	-0.01 (7/9)
	SE2	*0.40* (26/73)
Appeal	DE1	*0.42* (25/17)
	SE2	*0.40* (5/20)
Argumentative	DE1	0.32 (36/91)
	DE2	**0.63** (57/29)
	SE2	*0.58* (160/282)
Bio	DE1	0.28 (16/38)
	DE2	0.10 (3/33)
	SE2	**0.68** (20/18)
Fiction	DE1	**0.88** (45/57)
	DE2	0.00 (0/11)
	SE2	**0.75** (12/20)
Guidance	DE1	**0.88** (327/290)
	DE2	0.00 (0/59)
	SE2	**0.74** (452/423)
Informational	DE1	*0.43* (146/106)
	DE2	*0.46* (402/208)
	SE2	0.35 (418/433)
Interactive	DE2	0.00 (0/10)
Invitation	DE1	*0.46* (23/32)
	DE2	-0.01 (16/7)
	SE2	0.57 (191/134)
Legal	DE1	*0.50* (104/125)
	DE2	0.00 (0/5)
	SE2	0.25 (34/36)
News	DE1	0.21 (9/37)
	DE2	0.24 (64/89)
	SE2	-0.01 (12/73)
PO_report	DE1	0.08 (106/5)
	DE2	**0.71** (38/55)
	SE2	**0.86** (1204/1241)
QA	DE1	**0.97** (72/74)
	DE2	**0.66** (22/11)
	SE2	0.00 (69/0)

Table 5: IAA across Genre, Language, and Group. Counts (A1/A2) for each group are shown in brackets. Substantial agreement for Cohen's κ is highlighted with bold, moderate agreement – with cursive.

For the Swedish material, substantial agreement was observed for several genres, including patient organization reports (PO_report)—the largest and most characteristic genre of Swedish patient organization periodicals—as well as Guidance, Fiction, Biographical, Advertisement, and Academic texts. In the German material, similarly strong agreement was found for QA, PO_report, Guidance, Fiction, Argumentative, and Advertisement genres, although the agreement was not always consistent across groups.

In some cases, κ values dropped to zero, typically when one annotator assigned multiple instances to a given genre while the other classified all of them differently. We interpret these discrepancies as stemming from differences in annotators' weighting of contextual versus local information — that is, whether they prioritized the broader page context or the paragraph's own content when determining the genre.

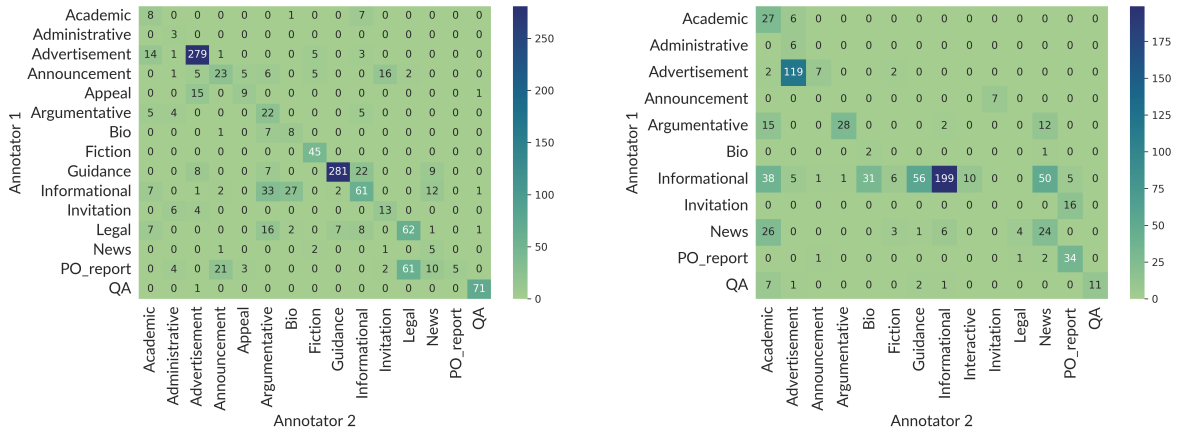
To better understand how specific genres were overlapping within each group's annotations, we examined the corresponding confusion matrices. In particular, for Swedish Group 2 (Fig. 2), the same paragraphs were frequently annotated as Advertisement by the first Annotator in pair (A1) and as Informational texts by the second one (A2). This appears to result directly from differing interpretations of local versus global context on the page. Advertisements in these periodicals often include detailed descriptions of goods, drugs, or medical equipment, which may or may not employ explicitly promotional language. Similar overlap can be observed between the Argumentative and Guidance genres, where factual or descriptive passages are often embedded within argumentative reasoning or instructional explanations.

In the German groups (Fig. 1), a similar pattern can be observed for the Informational genre. In both groups, A1 consistently assigned the label Informational, whereas A2 distributed these same paragraphs across several other genres—Argumentative and Biographical in Group 1, and Academic, Biographical, Guidance, and News in Group 2. An additional overlap in Group 1 occurred between Academic and Advertisement, likely because some advertisements include descriptions of research studies. Depending on how much contextual information the annotators considered, these could be interpreted either as research-related content or as part of an advertisement. In Group 2, Academic also overlapped with Argumentative, Informational, and News. This pattern may reflect A2's stronger reliance on lexical cues, as scientific terminology frequently appears across multiple genres in these periodicals, even though the underlying communicative function can differ substantially.

6.2. Data Availability

The annotated dataset is available for research purposes upon request. A link to the dataset's Hugging Face page is provided in the project's GitHub repository[8]. The repository also contains access

[8] https://github.com/ActDisease/genre-annotation-project



(a) Group 1.

(b) Group 2.

Figure 1: IAA in German annotation groups.

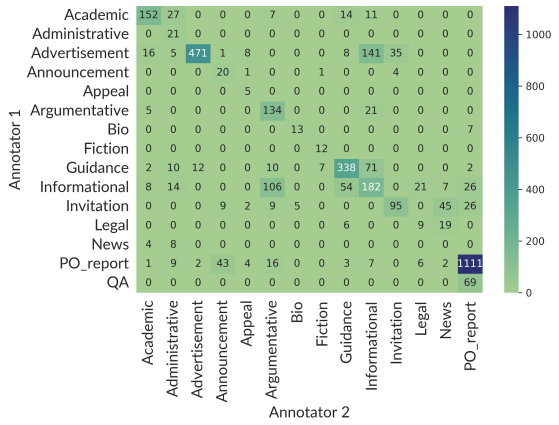


Figure 2: IAA in Group 2, Swedish Language

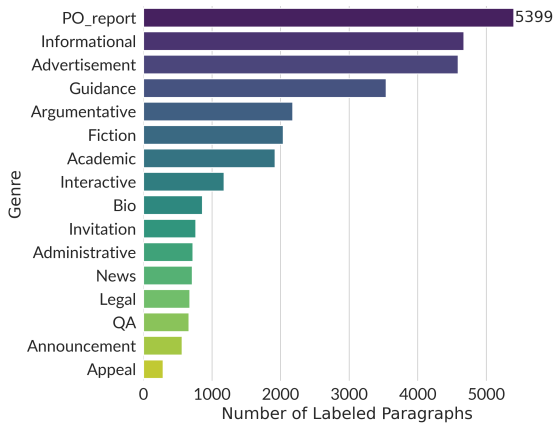


Figure 3: Label distribution in the Final Dataset

to the project's Notion workspace, which includes additional documentation and project materials.

7. Re-Annotation

A new round of annotation of the Swedish material was conducted over several months to provide a second opinion on the labels assigned by Annotator No.1 (see Tab. 1). This iteration involved two annotators: a project member and computational linguist familiar with the material and its genres, and a PhD student in the History of Science and Ideas. Both followed the same annotation guidelines within the Notion framework. The PhD student received the same training as previous annotators, including instruction on the annotation file, guidelines, and example cases.

The re-annotated sample comprised 996 paragraphs: 214 from Diabetes and 782 from Status. Across the three annotators (including Annotator No.1), Krippendorff's α was 0.82, with a percentage agreement of 82%.

Language	Periodical	Group 1	Group 2
German	Der Allergiker	3 (382), 4 (353)	5 (3405), 2 (816)
	Diabetiker Journal	3 (4061), 4 (951)	—
Swedish	Diabetes	1 (9376), 8 (214), 9 (214)	6 (4505), 7 (3788)
	Status	1 (986), 8 (782), 9 (782)	—

Table 6: Distribution of annotated paragraphs by periodical and group. Group cells show Annotator number and Size in round brackets.

8. Final Dataset Description

The distribution of genres in the final dataset is displayed in Fig. 3. The majority of annotated paragraphs belong to the genre "PO_report" (patient organization report), the most frequent genre in the Diabetes periodical and a defining feature of patient

organization publications. These reports combine narrative accounts, patient experiences, and informational content, reflecting the communicative goal of such periodicals — to inform and engage the patient community. Other common genres include Informational, Advertisement, and Guidance texts, which together characterize the patient periodical domain. Through these genres, editors and healthcare stakeholders communicate educational and promotional content, fostering interaction with their readership.

Tab. 6 summarizes the number of annotated paragraphs per periodical and language for each annotator group. Following the re-annotation process, we achieved near-complete coverage of the Status periodical sample with three annotators. Status is particularly rich in entertainment genres, with Fiction representing 45% and Interactive content 93% of such texts. Because fiction often overlaps with *biographical patient narratives* and domain-specific language, re-annotation was especially valuable for ensuring double coverage of this complex genre and improving annotation reliability.

9. Agreement Analysis

To further understand what contributes to annotator agreement, we analyze the relationship between paragraph- and genre-level features and annotator agreement.

Paragraph-level analysis To investigate whether overlapping annotations are related in terms of readability, lexical, and structural complexity, we compute a range of linguistic features for each paragraph using `spaCy`[9] and `textstat`[10]. The extracted features include measures of sentence and word structure (average sentence length, average word length), lexical diversity (type–token ratio), informational content (lexical density), and a composite readability score (Wiener–Sachtextformel (Bamberger and Rabin, 1984) for German and LIX (Björnsson, 1968; Falkenjack et al., 2013) for Swedish).

To capture syntactic complexity, we extract the average dependency tree height and average dependency distance between heads and dependents. We further calculate part-of-speech (PoS) ratios to quantify the distribution of linguistic categories, including nouns, verbs, adjectives, adverbs, adpositions, pronouns, subordinating and coordinating conjunctions, particles, auxiliaries, numerals, proper nouns, and punctuation marks. Together, these features characterize lexical choice, syntactic structure, and stylistic variation across texts.

For each feature, we fit a logistic regression model predicting annotator agreement (1) versus disagreement (0). The predictive performance of each feature is evaluated using the *Area Under the Curve* (AUC)

Genre-level analysis At the genre level, we use three types of representations for each genre and annotator: (1) a vector of averaged linguistic features, as described above, (2) a semantic embedding vector, and (3) a topic vector. For the semantic vectors, all paragraphs are embedded using the `Qwen3-4B-embeddings` model[11]. Topic representations are derived using `BERTopic` (Grootendorst, 2022), applying UMAP (McInnes et al., 2020) and HDBSCAN (McInnes et al., 2017) clustering of the Qwen 4B embeddings to identify latent topic clusters.

For the first two representations mean pairwise cosine similarity is computed between each pair of genres across annotators (using `sklearn`).

Topic overlap between genres of each pair of annotators is then quantified as follows based on the topic vectors. For each pair of genres assigned by the two annotators, we extract the corresponding sets of unique topics. We then compute the overlap as a percentage by dividing the number of shared topics (the intersection) by the total number of unique topics across both sets.

For each representation, both the agreement matrix and the resulting similarity matrix are flattened into 1D arrays to compute Spearman's rank correlations. This calculation is conducted on two levels. A full-matrix correlation is first calculated to capture general trends across all annotator decisions. Subsequently, to isolate and reflect relationships specifically associated with annotator disagreement, the matrices are filtered by removing their diagonal elements (which represent perfect agreement) before computing the off-diagonal correlations.

Association between topic and agreement To further examine whether agreement patterns depend on topic distribution, we apply a Chi-squared test, with the strength of the association quantified using Cramér's V (Agresti, 2007). Higher values indicate a stronger alignment between topic and annotator agreement.

Results Tab. 7 summarizes the strength and significance of the association between topic distribution and annotator agreement (Cramér's V), alongside the correlations between genre-level similarity

[9] https://spacy.io
[10] https://pypi.org/project/textstat/
[11] https://huggingface.co/Qwen/Qwen3-Embedding-4B

Group	V	Sem. ρ	Struct. ρ	Topic ρ
DE1	0.70^{**}	0.19^{**}	0.33^{**}	0.47^{**}
DE2	0.59^{**}	0.21^{**}	0.47^{**}	0.44^{**}
SE1	0.71^{**}	0.50^{**}	0.52^{**}	0.71^{**}
SE2	0.57^{**}	0.01	0.50^{**}	0.61^{**}

Table 7: Association between topic distribution and annotator agreement (V–Cramér's V), and the correlations between similarity measures and disagreement. $(^{**}) - p < 0.01$

Group	Top Features (AUC)
DE1	AvgTreeHeight (.67), AvgSentLen (.67), VerbRatio (.65), AvgDepDist (.65), AuxRatio (.64)
DE2	SconjRatio (.54), AdpRatio (.54)
SE1	PunctRatio (.80), AvgTreeHeight (.79), AvgSentLen (.78), Lix (.78), AvgDepDist (.75)
SE2	PunctRatio (.58), AdvRatio (.58), VerbRatio (.57), PronRatio (.57), AuxRatio (.56)

Table 8: Top predicting linguistic features (AUC $>$ 0.55 at p-value $<$ 0.01).

measures and disagreement (off-diagonal calculation).

Across all groups, topic overlap shows the strongest and most consistent positive correlation with agreement (Cramér's V ranging from 0.57 to 0.71), suggesting that annotators are most likely to agree on paragraphs belonging to similar topical clusters. Structural similarity, derived from aggregated linguistic feature vectors, demonstrates a moderate but significant association with agreement, particularly in the Swedish groups (0.50). Semantic similarity, while generally weaker, remains statistically significant in most cases, indicating some alignment in interpretive focus across annotators.

Tab. 8 presents the top paragraph-level linguistic features identified as significant discriminators of agreement, ranked by their AUC scores. Syntactic indicators–dependency tree height, sentence length, and dependency distance–emerged as the most reliable discriminative features, achieving AUC values between 0.65 and 0.80. In contrast, lexical features (e.g., type–token ratio, lexical density) contributed less discriminative power. The strongest effects were observed in Swedish Group 1, where multiple syntactic measures reached AUC $>$ 0.75, underscoring the role of structural regularity and readability in driving consistent annotation behavior.

These results suggest that agreement between annotators is more strongly linked to structural and topical coherence than to purely lexical similarity, reflecting shared judgments about text organization and content domain rather than word-level overlap.

10. Conclusion

We introduce a new OCR-processed dataset of German and Swedish medical periodical texts annotated with textual genre, retaining individual annotations. Genre annotation of these texts is challenging due to the their historical, medical, and journalistic nature. High agreement is achieved for both languages in Advertisement, Fiction, Guidance, and PO_report. Disagreements often arise from differences in how annotators interpret the context surrounding paragraphs. The analysis of correlation of different metrics with agreement patterns shows that topic distributions correlate with it significantly, meaning that unknown topics are likely to raise doubts on genre assignment. Structural complexity also significantly impacts agreement, which is especially noticeable in the largest double-annotated samples for each language.

As observed in previous work (Peter and Lauf, 2002), linguistic and cultural background influence annotation reliability. In future work, this dataset will be used to examine how background factors and category definitions affect reliability, as well as to explore the role of linguistic factors, the hybrid nature of genres, and differences between paragraph- and column-level interpretations of communicative purpose. A genre classifier will be trained to investigate patterns in communicative strategies.

11. Ethical Statement

This study involves the annotation of historical OCR-processed German and Swedish texts by multiple human annotators. All annotators participated voluntarily and were fully informed about the annotation objectives, and potential sensitivities related to medical information. They were also made aware of the restrictions on the sharing and use of these materials in accordance with the project's ethical framework. The periodicals used were publicly available historical publications distributed in their respective countries; no confidential or unpublished data were processed. Annotators were compensated with an hourly wage in accordance with university regulations.

12. Acknowledgements

The work on genre classification of historical medical periodicals forms part of the ActDisease project, funded by the European Research Council (ERC, ActDisease, ERC-2021-STG 10104099). The views and opinions expressed are those of the authors alone and do not necessarily reflect the positions of the European Union or the European Research Council. Neither the European Union nor the funding agency can be held responsible for

them. We are grateful to the ActDisease historians for their contributions to the analysis and definition of genres in historical medical periodicals.

The annotation project was also supported by the Swedish national research infrastructure Språkbanken, with joint funding from the Swedish Research Council (2018–2028; grants 2017-00626 and 2023-00161) and the ten participating partner institutions.

We extend our sincere thanks to the annotators who generously contributed to this project: Sebastian Kreft, Yanitsa Stoykova, Jannis Schuh, Ida Nilsson, Nele Popp, Hanne Johansson, Anne Andersen, and Majken Cederman.

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725, and the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

13. Bibliographical References

Alan Agresti. 2007. *An Introduction to Categorical Data Analysis*, 2nd edition. John Wiley & Sons.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Richard Bamberger and Annette T. Rabin. 1984. New approaches to readability: Austrian research. *The Reading Teacher*, 37(6):512–519.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Aysenur Bilgin, Erik Tjong Kim Sang, Kim Smeenk, Laura Hollink, Jacco van Ossenbruggen, Frank Harbers, and Marcel Broersma. 2018. Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 486–496.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Lärarbiblioteket. Liber, Stockholm.

Tommaso Caselli, R. Sprugnoli, and Giovanni Moretti. 2022. Identifying communicative functions in discourse with content types. *Language Resources and Evaluation*, 56:417–450.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40, Oslo, Norway. Linköping University Electronic Press, Sweden.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint 2203.05794.

Frank Harbers. 2014. *Between personal experience and detached information: The development of reporting and the reportage in Great Britain, the Netherlands and France, 1880-2005*. Ph.D. thesis, University of Groningen.

N. Ide and J. Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Number v. 1 in Handbook of Linguistic Annotation. Springer.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain. Association for Computational Linguistics.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey: Automatic genre identification: a survey. *Lang. Resour. Eval.*, 59(1):537–570.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation

and projection for dimension reduction. arXiv preprint 1802.03426.

Jochen Peter and Edmund Lauf. 2002. Reliability in cross-national content analysis. *Journalism & Mass Communication Quarterly*, 79(4):815–832.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13:65–95.

Karin Stahel, Irenie How, Lauren Millar, Luis Paterson, Daniel Steel, and Kaspar Middendorf. 2025. A bit of this, a bit of that: Building a genre and topic annotated dataset of historical newspaper articles with soft labels and confidence scores. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 377–392, Albuquerque, USA. Association for Computational Linguistics.

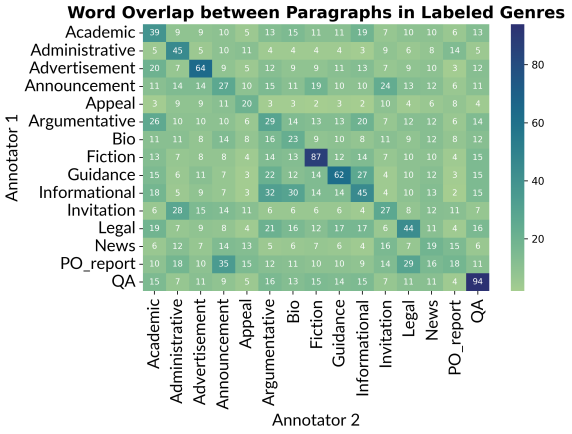
A. Example of the Annotated Dataset

annotator_id	part_no	language	fname	periodical	year	nr	page	paragraph	genre	other	ocr_errors	opinionated	dialogue
3	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Die Bezeichnung „Heufieber" oc	Argumentative		0.0	0.0	0.0
3	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Die Jahresberichte stellten an d	Administrative		0.0	0.0	0.0
3	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Seit Bestehen unseres Bundes l	Administrative		0.0	0.0	0.0
3	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Waren Blütezeit und Beschwerd	Administrative		0.0	0.0	0.0
4	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Die Bezeichnung „Heufieber" oc	Informational		0.0	1.0	0.0
4	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Die Jahresberichte stellten an d	Administrative		0.0	1.0	0.0
4	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Seit Bestehen unseres Bundes l	Administrative		0.0	0.0	0.0
4	1	deu	Der_Aller	xml_der_allergiker	1959	nr001	0	Waren Blütezeit und Beschwerd	Administrative		0.0	1.0	0.0

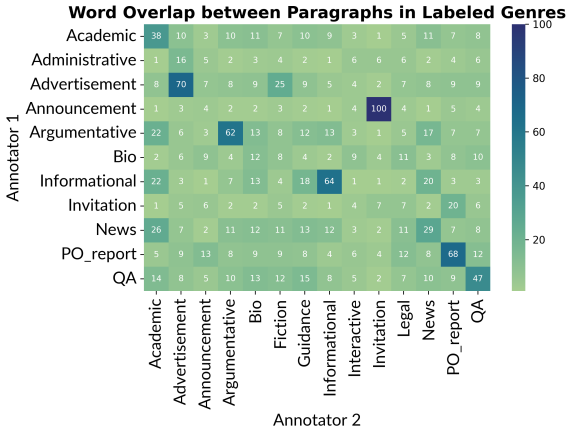
Figure 4: Annotated dataset example

B. Word Overlap in Annotation Groups

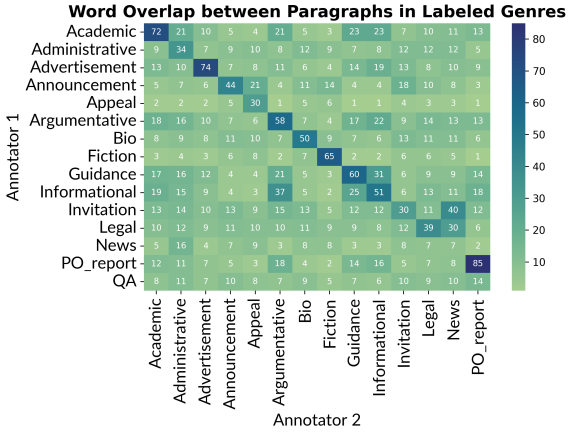
To additionally quantify lexical similarity between genres, word overlap was calculated using the Jaccard similarity coefficient. For every pair of genres assigned by Annotator 1 and Annotator 2, all unique tokens from the corresponding texts were aggregated into two distinct sets. The overlap was then computed by dividing the number of shared words (the intersection) by the total number of unique words across both sets. The final score was converted to a percentage to populate a genre-by-genre similarity matrix:



(a) German Group 1



(b) German Group 2



(a) Swedish Group 2