Genre Annotation of Web Corpora: Scheme and Issues

Vít Suchomel[1,2]

[1] Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz
https://nlp.fi.muni.cz/en/
[2] Lexical Computing, Brno, Czech Republic
https://www.lexicalcomputing.com/

Abstract. Unlike traditional corpora made from printed media in the past decades, sources of web corpora are not categorised and described well, thus making it difficult to control the content of the corpus.
This paper presents an attempt to classify genres in a large English web corpus through supervised learning. A set of genres suitable for web corpora users is defined based on a research of related work. A genre annotation scheme with active learning rounds is introduced. A collection of web pages representing various genres that was created for this task and a scheme of consequent human annotation of the data set is described.
Measuring the inter-annotator agreement revealed that either the problem may not be well defined, or that our expectations concerning the precision and recall of the classifier cannot be met.
Eventually, the project was postponed at that point. Possible solutions of the issue are discussed at the end of the paper.

Keywords: Corpus annotation, Inter-annotator agreement, Text genre, Web corpora

1 Motivation: Richer Web Corpora

A corpus is a special collection of textual material collected according to a certain set of criteria. In statistical natural language processing one needs a large amount of language use data situated within its textual context. The text corpora are one of the main requirements for statistical NLP research. [17, pp. 5, 6, 117, 119]

The field of linguistics greatly benefits from the evidence of language phenomena of interest one can found in large text corpora. In particular such data source is essential for various subfields of computational linguistics such as lexicography, machine translation, language learning, text generation.

Understanding the content of a web corpus is important. Traditional corpora were designed for particular use and compiled from deliberately selected sources of good quality:

– British National Corpus consists of ca. 10 % spoken data and ca. 90 % written component further divided by domain, genre, level and date. [15]
– Czech National Corpus SYN, 2015 edition – as reported in [12] – consists of 1/3 fiction, 1/3 non-fiction and 1/3 newspapers and magazines. These sources are further sub-classified in a hierarchy with 40 classes on the lowest level. Text types in the corpus are purposefully balanced, e.g. there is the same amount of documents in these second level sub-classes of class non-fiction: humanities, social sciences, natural sciences and technical sciences, each of them taking 7 % of the whole corpus.
– Slovak National Corpus is reported to contain published texts consisting of 71.1 % journalism, 15.4 % arts, 8.5 % technical documents, 5.0 % other texts.[3]
– Estonian National Corpus is a mix of texts from traditional media and web pages with subcorpora from Estonian Reference Corpus 1990-2008, Estonian web from 2013, 2017 and 2019, Estonian Wikipedia from 2017 and 2019, Estonian Open Access Journals (DOAJ). [9] There is rich metadata in the Reference corpus part and the Wikipedia subcorpus because of the nature of their sources.
– Wikipedia based corpora can provide language varieties, topics (Wikipedia portals and categories) and links to related pages. Obviously, there is just the informative or encyclopedic genre present in Wikipedia pages.

Such precise selection of nice texts is hardly possible in the case of large web corpora. Nevertheless, those who build web corpora should provide information to enable researchers using these corpora know their sources (compare to the appeal in [10] where corpus users are held responsible for knowing their sources). We should study and categorize what is being downloaded from the web.

From our point of view, it is desirable to provide web corpora with rich annotation such as language varieties (e.g. British, American, Australian or Indian English), topics (such as sports, health, business, culture), genres (informative, encyclopedic, instructive, appellative, narrative), registers (formal, informal, professional) and other text types.

Since these categories do not come with the data, it is possible to annotate web documents using a supervised classification approach. Furthermore, the distribution of these text attributes in a web corpus can help the users not only to know what is 'inside'. In addition to that, it enables the users to work with just a selection of subcorpora based on the text types.

2 Genre Selection and Reliability of Classification

A dictionary definition of genre is 'A particular style or category of works of art; esp. a type of literary work characterised by a particular form, style, or purpose.'[4]

[3] https://korpus.sk/structure1.html, accessed in January 2020.
[4] The second edition of Oxford English Dictionary (1989). Accessed online at https://www.oed.com/oed2/00093719 in April 2020.

[18] state genre is 'A set of conventions (regularities) that transcend individual texts, helping humans to identify the communicative purpose and the context underlying a document.'

To add a perspective of text corpus users who do language research, build dictionaries, or e.g. produce language models for writing prediction – adding information about genre to corpus texts allows them to know more about the composition of the corpus and enables them to use subcorpora limited to a particular genre.

[19] lists reasons for determining genres of web documents for information retrieval. These reasons can be applied to corpus linguistics and NLP too:

– 'People normally store and retrieve documents by genre.'
– 'Genre is a compact way to describe a document.'
– 'There is a need for non-topical search descriptors.'
– 'Many traditional genres have migrated to the web.'
– There are unique genres on the web. 'Some of the most popular tags for web pages on the social tagging site `delicio.us` are genre labels, such as blog, howto, tutorial, news and research.'

[1, p. 121] uses The Lancaster-Oslo/Bergen Corpus of British English (LOB) and The London-Lund Corpus of Spoken English. Six dimensions based on lexical, syntactic and semantic attributes of text are described, e.g. Narrative vs. Non-Narrative Concerns or Abstract vs. Non-Abstract Information.

Indeed, genres can be described by linguistic properties of text. [2] uses the term 'linguistic markers' and lists positive features of e.g. discerning 'narrative versus non-narrative concerns' – 'past-tense verbs, 3rd person pronouns, perfect-aspect verbs, public verbs, synthetic negation, present-participial clauses' – and complementary features – 'present-tense verbs, attributive adjectives'.

Biber used rules based on scores of linguistic features. Machine learning techniques are preferred by recent research. Our understanding is a genre is determined by the style of writing, content words are only supporting evidence – unlike topics that are determined by content words. Therefore it is the style that is key in assessing the genre, content words are only secondary. It is necessary to add linguistic features such as the verb tense in the features for training a classifier.

Well known corpora Brown Corpus and LOB consist of a priori determined number of texts bearing signs of the following genres and topics: Press: reportage; Press: editorial; Press: reviews; Religion; Skills, trades and hobbies; Popular lore; Belles lettres, biography, essays; Learned and scientific writings; General fiction; Mystery and detective fiction; Science fiction; Adventure and western fiction; Romance and love story; Humour; Miscellaneous.

Unlike corpora traditionally constructed from selected sources with known text types, we need to classify documents that come from sources without a determined text type and that are already in the corpus.

In the early ears of the internet, [5] dealt with the genre of web sites. The set of genres to recognise was derived from a poll of internet users. The result set was: Personal homepages; Public or commercial homepages; Searchable indices;

Journalistic materials; Reports (scientific, legal); Other running text; FAQs; Link Collections; Other listings and tables; Asynchronous multi-party correspondence (discussions, Usenet News); Error Messages. Despite too internet oriented and somewhat old, categories such as Personal homepages, Public or commercial homepages and FAQs are useful for our purpose.

While [19] claimed ‘People can recognize the genre of digital documents’, [22] stated the opposite – that humans are incapable of determining the genre of a web page consistently and that web pages can have multiple genres and may not resemble sample single genre texts.

A research of papers on web genres by [3] revealed serious difficulties of defining and determining web genres: ‘Unfortunately, our review of the literature reveals a lack of consensus about the Web genre taxonomy on which to base such systems. Furthermore, our review of reported efforts to develop such taxonomies suggests that consensus is unlikely. Rather, we argue that these issues actually resist resolution because the acceptance of potential answers depends on a researcher’s epistemological and ontological orientation’. Yet they stress ‘a continuing and, indeed, growing need for understanding a document’s genre’.

In our work, we are determined to identify genres useful for text corpora users. That should reduce the number of possibilities and approaches at least a little.

[22] claimed ‘An inherent problem of Web genre classification is that even humans are not able to consistently specify the genre of a given page.’ Such cautious approach resulted in a relatively small count of genres that were identified: Help; article; discussion; shop; private portrayal; non-private portrayal; link collection; download. A support vector machines classifier was trained on 800 HTML pages leading to an average classification performance of 70 %.

From our point of view, Dewe [5], zu Eissen and Stein [22] and Crowston [3] propose too web oriented labels. We believe corpus linguists and lexicographers expect separate classes like Reporting, Information, Legal, Narrative which are not discerned by these works.

The audience of British National Corpus (BNC) is close to ours. [14] categorised documents in BNC into 46 genres of written text and 24 genres of spoken text. Some documents can have multiple genres assigned.

Such fine grained categorization is not possible in the case of web corpora for the reasons already mentioned.

[11] constructed a corpus from four sources naturally consisting of different genres: Conversation, newspaper, fiction and the web. (The point was a dictionary based on that corpus represented the use of words in those genres better than if it had been based on a single genre corpus.)

On the contrary, we would like to keep selected web genres separate rather than merging them into a single label.

[4] identified seven genres in academic texts on the web: Instructions, hard news, legal, commercial presentation, science/technology, information, narrative. Although these categories represent some genres on the web, we need to cover more than academic texts.

Following Biber, [21] worked with 'Functional Text Dimensions' and focused on large web corpora. 18 genres were identified: Argumentative, Emotive, Fictive, Flippant, Informal, Instructive, Hard news, Legal, Personal, Commercial presentation, Ideological presentation, Science and technology, Specialist texts, Information or encyclopedic, Evaluation, Dialogue, Poetic, Appellative. A subset containing 12 of these genres was also defined.

We decided to start from the Sharoff's list of 12 genres. As noted above, finding genres that can be reliably identified is a difficult task. We understand borders between genres are not sharp so other parameters, namely the very definition of genres, have to be adapted.

Therefore we decided to measure the agreement of human classification of genres of web documents and merge classes until the inter-annotator agreement is sufficient. The agreement can be increased by decreasing the granularity of genres to discern. That is why we call our approach Agreement driven.

3 Experiment Setup

The set of 12 genres defined by [21] is: Argumentative, Fictive, Instructive, Hard news, Legal, Personal, Commercial presentation, Ideological presentation, Science and technology, Information or encyclopedic, Evaluation, Appellative.

Non-text (machine generated text and other web spam) was a 13th genre added by us to enable us to use the data for learning a non-text classifier for an experiment not covered by this paper.[5]

We aimed to determine genres of web pages in a large English web corpus by training a supervised classifier on human annotated texts. A new collection of web pages was created for this purpose. Texts were added in the collection in four subsets according to our evaluation of the task in several stages. The sources of the collection are summarised in Table 1.

Texts in Subset 1 come from various web and non-web sources: fiction, UN documents, TED talks, etc. (originally annotated as reported in [7]) and randomly selected texts from UKWaC [6] (originally annotated as reported in [21]).

Subset 2 contains another set of UKWaC texts selected as those with the most certain annotations by applying the classifier from [4].

Texts in Subset 3 were obtained from enTenTen13 [8]: Documents in the corpus were classified by a FastText classifier trained on documents in Subsets 1 & 2. Web pages with the lowest certainty of the classifier (measured by the entropy of the probability distribution of labels given by FastText for a particular document) were used.

We did four rounds of manual annotation of texts from the collection. A group of students and academics at the University of Leeds was instructed and supervised by Serge Sharoff. Another group of students and academics at Masaryk University was instructed and supervised by the author of this paper.

[5] Since this paper is about genres in web corpora rather than web spam removal, genre Non-text is not included in the results unless explicitly mentioned.

Table 1. Sources of the collection of texts used in our experiments. Different subsets (**S**) were added in different times (starting with Subset 1). UKWaC, enTenTen13, enTenTen15 and enTenTen18 are English web corpora from 2007, 2013, 2015 and 2018, respectively.

S	Description	Author	Count
1	Selection of web and non-web texts	Sharoff & others	448
	Selection of non-text from enTenTen15	Suchomel	123
2	Most certain texts from UKWaC	Sharoff	456
	Web search for underrepresented genres	Suchomel	198
3	Least certain texts from enTenTen13	Suchomel	405
4	enTenTen18 random texts	Suchomel	344
Total count of documents in the collection			1974

The inter-annotator agreement (IAA) was measured after each round to see if the genre definition was understood well and to decide how to improve. Sample documents to explain genre differences to annotators were created. Multiple labels were allowed for documents showing strong signs of more genres.

We also did a round of active learning: According to the idea of a technique called Uncertainty sampling [16], annotating samples where the classifier is the least certain should efficiently contribute to training the classifier, e.g. requiring a lower number of costly annotated samples than a random selection.

'The basic premise is that the learner can avoid querying the instances it is already confident about, and focus its attention instead on the unlabeled instances it finds confusing.' [20, p. 11]

In our experiment, a classifier was trained on texts annotated at that time using FastText. Documents with the highest entropy of the probability distribution of labels provided by FastText – i.e. cases where the classifier was most unsure – were selected for the next round of the annotation.

Since the initial IAA was below our expectation, we tried merging labels and omitting least successful classes getting to '6 classes + non-text'.

Unfortunately, an evaluation showed that did not help enough so we decided to start over with the following set of categories (still based on Sharoff's genres from [21] a lot).

1. Information – subcategorised to
   (a) Promotion (covering both commercial and ideological presentation from Sharoff's list): Promotion of a product, service, political movement, party, religious faith. Examples: An advert, a product/service page, a political manifesto.
   (b) Academic: Research. Example: A research paper or any text written using the academic style.

   (c) Review: Evaluation of a specific entity by endorsing or criticising it. Example: A product review endorsing or criticising the product.
   (d) Other: Any informative text not from a category above.
2. Story telling (a better name for both Fiction and Narrative): A description of events (real or fictional, usually in the order they followed), often informal, can be in the first person. Examples: Fiction, narrative blogs.
3. Instructions: Teaching the reader how something works. The imperative is frequent. 2nd person pronouns may be frequent. Examples: Howtos, FAQs, instructions to fill a web form.
4. News: Informative report of events recent (or coming in the near future) at the time of writing (not a discussion or a general state of affairs). Usually the formal style, set to a particular place and time. May quote sources of information. Example: A newswire, diary-like blogs.
5. Legal. Examples: A contract, a set of regulations, a software licence.
6. Discussion: A written communication of participants of a discussion. Usually multiple authors. Can be personal, informal style. Examples: Web forums, discussions, product comments.
7. Unsure or too short: Indicating the annotator was unable to determine the genre with confidence.

Four subcategories of Information were meant to be used separately or merged together for the final classifier depending on the IAA and classifier performance.

The fourth subset of the collection was created from enTenTen18 at this moment to introduce more contemporary web documents to the collection. A way for annotators to mark documents too short to reach an agreement was implemented to reduce noise in the training data.

A web application providing the annotation interface was made by the author of this paper. The tool consist of a Python script and a Sqlite database backend and an HTML & JavaScript frontend making asynchronous requests to the backend. Screenshots of the application can be seen on Figure 1 and Figure 2. The functionality of the interface includes:

– Displaying the plaintext to annotate with a link to the original web page.
– Compact definitions of genres with examples from the annotation manual are shown. A short description of the procedure helps annotators to remember key concepts, e.g. that the style of writing is more important than the topic.
– Metainformation such as annotator's nickname, number of documents already annotated and remaining to annotate in the current round is displayed.
– Multiple genres can be marked for each document.
– Texts too short to determine a genre or undetermined for any reason can be marked.
– There are four labels to indicate the presence of the features of a particular genre in the document: None (the default, initially selected for each genre), Somewhat, Partly, Strongly.

- The plaintext is split to paragraphs.
- The application allows marking paragraphs containing signs of minor genres different from the genre of the majority of the text. This function can help to reduce the noise by removing non-text paragraphs and minor genre paragraphs from the training data.
- A review mode allowing the supervisor to review annotations of others after a training round.
- Anyone can be enabled to use the review mode to see common mistakes or example documents after a training round.

After reviewing the initial round of annotations, we decided to account in only 'Strongly' labels. The annotators did not know that. The purpose of other, weaker labels was just to help humans resist the temptation to choose 'Strongly' if not perfectly sure (and not indicating 'Unsure or too short' too).

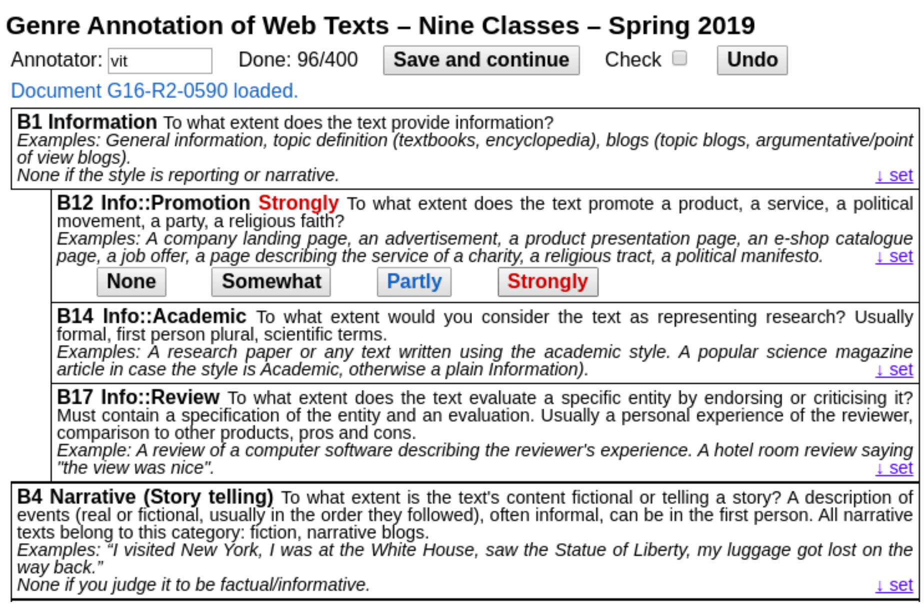

Fig. 1. Text type annotation interface – web application in a browser – the left side of the screen. Information about the annotation process can be seen at the top. Genres with a brief description and examples follow. Class 'Information::Promotion' is labelled as strongly present in this case. Buttons for weaker presence of genre markers (Partly, Somewhat, None) can be clicked to change the annotation.

The final experiment was continued by hiring six university students proficient in English at level B2 or C1 who received a two hours' training. Then, each of annotators worked through a training round of 45 documents. All annotations

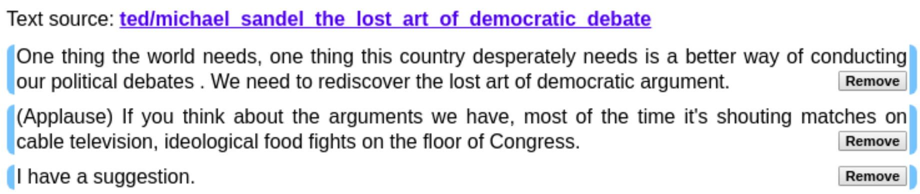

Fig. 2. Text type annotation interface – web application in a browser – the right side of the screen. The title of the document with a link leading to the original source is located at the top. The plaintext split to paragraphs can be seen below. Both sides of each paragraph are coloured to visualise separate paragraphs. A paragraph can be suggested for removal from the document (to make the result training data less noisy) by clicking the respective button. The source of the text in this screenshot: https://www.ted.com/talks/michael_sandel_the_lost_art_of_democratic_debate/transcript.

were checked by us, issues were made clear and everyone had to review frequent mistakes.

A screenshot of the use of the review mode can be seen on Figure 3. Since the agreement was poor in this instance, the case was explained to the annotators and the document was put to a review for everyone.

Annotator	B1	B12	B14	B17	B4	B7	B8	B9	B11	B22	B98	B99	Date	Time
anicka	0	2	0	0	0	0	0	0	0	0	0	0	2019-04-03	59
danca	0	0	0	0	0	2	0	0	0	0	0	0	2019-04-05	88
ivca	0	2	0	0	0	0	0	0	0	0	0	0	2019-04-04	201
kacka	0	2	0	0	0	0	0	0	0	0	0	0	2019-04-07	116
marky	0	0	0	0	0	2	0	0	0	0	0	0	2019-04-04	110
mates	0	0	0	0	0	0	0	2	0	0	0	0	2019-04-05	71
vit	0	1	0	0	0	0	0	2	0	0	0	0	2019-04-05	99

Fig. 3. Text type annotation interface in the review mode after the training round – as seen by the author of this paper who trained six other annotators. Labels, coded by identifiers in columns B1 to B99, assigned to a single document by each annotator are shown. Values 'Strong', 'Partially' and 'None' are coded by 2, 1, $^1/_2$ and 0, respectively. Time in seconds spent by annotating the document by each annotator can be seen in the rightmost column.

After the training round, two consecutive rounds of annotation were made. Several rounds of active learning were planned after that, i.e. training a classifier to find the most unsure documents and annotating those texts. However, we were not satisfied with the IAA at that time. Further evaluations were made to find if changing some parameters of the experiment would help.

The full definition of genres and instructions for annotators can be found in the Appendix.

4 Inter-annotator Agreement

Our goal was to reach the following level of inter-annotator agreement:
- Pairwise Jaccard's similarity ≥ 0.8. ... The similarity of sets of labels assigned by each pair of annotators is measured.
- Krippendorff's alpha[13] ≥ 0.67.[6]

Since multiple labels for a single sample were allowed, a text annotation was a set of labels in our case. We selected the following metrics providing the similarity of a pair of sets.

1. Accuracy as a set similarity metric:

```
def accuracy(labels1, labels2):
    i = labels1.intersect(labels2)
    return (len(i) / len(labels1) +
        len(i) / len(labels2)) / 2.0
```

2. Jaccard's similarity:

```
def jaccard(labels1, labels2):
    i = len(labels1.intersect(labels2))
    u = len(labels1.union(labels2))
    return len(i) / len(u)
```

3. Nominal comparison which tests for an exact match. It is the default metric for Krippendorff's alpha for discrete labels.

An example to illustrate the difference of the metrics (assume A, B, C are class labels) follows:

$$Acc(\{A,B\},\{A,C\}) = \frac{\frac{|\{A,B\}\cap\{A,C\}|}{|\{A,B\}|} + \frac{|\{A,B\}\cap\{A,C\}|}{|\{A,C\}|}}{2} = \frac{\frac{1}{2}+\frac{1}{2}}{2} = \frac{1}{2}$$

$$Jaccard(\{A,B\},\{A,C\}) = \frac{|\{A,B\}\cap\{A,C\}|}{|\{A,B\}\cup\{A,C\}|} = \frac{1}{3}$$

$$Nominal(\{A,B\},\{A,C\}) = 0 \ (\{A,B\} \neq \{A,C\})$$

An overview of all experiments with IAA expressed as Accuracy, Jaccard's similarity and Krippendorff's alpha with set similarity metrics Accuracy, Jaccard's similarity and Nominal comparison is provided by Table 2.

The first four experiments led to the final setup with nine classes (in bold typeface in Table 2). All rows below were derived from that setup by not counting instances marked as unsure or not counting instances with multiple labels or by merging labels Information, Promotion, Academic and Review into a single label.

[6] Krippendorff wrote on the acceptable level of reliability expressed by the α: 'Rely only on variables with reliabilities above $\alpha = .800$. Consider variables with reliabilities between $\alpha = .667$ and $\alpha = .800$ only for drawing tentative conclusions.' [13, p. 241]

Table 2. Inter-annotator agreement of genre annotation of web documents for different experiment setups. **P** is the count of people annotating, **Data** refers to collection subsets, **N** is the count of documents, **A** is the average count of annotations per text. **Acc** is Accuracy, **Jac** is Jaccard's similarity, **K-Acc**, **K-Jac** and **K-Nom** stand for Krippendorff's alpha with the set similarity metric set to Accuracy, Jaccard's similarity and Nominal comparison, respectively. '6/9 genres' means that four of the nine labels were merged in a single label for the particular evaluation. 'No unsure' means annotations indicating the person was not sure were omitted. 'No multi' means annotations with multiple strong labels were omitted.

Experiment	P	Data	N	A	Acc	Jacc	K-Acc	K-Jac	K-Nom
12 genres + spam	7	1 & 2	77	3.66	0.527	0.507	0.444	0.428	0.401
12 genres + spam	4	1 & 2	149	3.30	0.495	0.484	0.449	0.438	0.417
6 genres + spam	4	1 to 3	50	4.00	0.660	0.653	0.557	0.550	0.534
6 genres + spam, no unsure	5	1 to 3	50	5.00	0.768	0.762	0.603	0.595	0.580
9 genres, training	7	All	45	7.00	0.600	0.585	0.491	0.477	0.449
9 genres – the base for ↓	**6**	**All**	**1,356**	**2.57**	**0.640**	**0.628**	**0.530**	**0.518**	**0.497**
9 genres, no unsure	6	All	1,342	2.51	0.670	0.658	0.562	0.550	0.528
9 genres, no unsure, no multi	6	All	1,340	2.43	0.676	0.676	0.570	0.570	0.570
6/9 genres	6	All	1,356	2.57	0.776	0.765	0.566	0.552	0.527
6/9 genres, no unsure	6	All	1,342	2.51	0.814	0.802	0.622	0.606	0.576
6/9 genres, no unsure, no multi	6	All	1,340	2.43	0.819	0.819	0.629	0.629	0.629

Figure 4 and Figure 5 show pair annotation matrices for experiments without unsure or multi-label samples accounted in. Each pair of annotations of the same sample by two annotators was accounted in a two dimensional matrix with each dimension representing labels given by the corresponding annotator, i.e. a sample was accounted in the row respective to the label given by the first annotator and the and column respective to the label given by the second annotator. Agreements are on the diagonal, disagreements are in other fields. The percentage of all pairs is shown.

9 genres	Info	Story	Instr	News	Legal	Disc	Prom	Acad	Rev
Information	**27.7%**	2.9%	2.9%	6.4%	0.8%	0.8%	9.0%	2.4%	1.3%
Story telling		**7.0%**	0.1%	0.2%	0.0%	0.1%	0.5%	0.0%	0.0%
Instructions			**4.4%**	0.0%	0.1%	0.2%	1.0%	0.0%	0.2%
News				**5.9%**	0.0%	0.1%	0.7%	0.1%	0.3%
Legal					**3.2%**	0.0%	0.2%	0.0%	0.0%
Discussion						**0.7%**	0.1%	0.0%	0.4%
Promotion							**14.5%**	0.1%	1.3%
Academic								**1.3%**	0.1%
Review									**2.9%**

Fig. 4. Pair annotation matrix for the setup with 9 genres, without unsure or multi-label samples. Percentage of all annotation pairs is shown.

6/9 genres	Info	Story	Instr	News	Legal	Disc
Information	**60.6%**	3.4%	4.1%	7.5%	1.0%	1.3%
Story telling		**7.0%**	0.1%	0.2%	0.0%	0.1%
Instructions			**4.4%**	0.0%	0.1%	0.2%
News				**5.9%**	0.0%	0.1%
Legal					**3.2%**	0.0%
Discussion						**0.7%**

Fig. 5. Pair annotation matrix for the setup with 6/9 genres, without unsure or multi-label samples. Percentage of all annotation pairs is shown.

It can be seen that Information was the class causing most disagreement. The reason may be that borders between other genres are more clear than the border of any genre with Information in our definition of genres.

The percentage of agreement for each class (i.e. the ratio of the value on the diagonal to the rest related to the label) is summarised in Table 3.

Table 3. Pair agreement summary for setups with 9 genres and 6/9 genres, without unsure or multi-label samples.

Pair agreement	**9 genres**	**6/9 genres**
Information	51.2 %	77.8 %
Story telling	64.8 %	64.8 %
Instructions	49.1 %	49.1 %
News	42.7 %	42.7 %
Legal	73.5 %	73.5 %
Discussion	31.7 %	31.7 %
Promotion	52.9 %	
Academic	32.4 %	
Review	44.1 %	

5 Dealing with a Low Agreement

To summarise, we tried the following to improve the inter-annotator agreement:

- The number of recognised genres was reduced.
- Multi-genre texts were omitted.
- Short texts (indicated by annotators) were omitted.
- Annotators were trained and their mistakes in the training round were explained thoroughly.
- Annotators could indicate they were not sure.
- Annotators were paid for time spent annotating rather than the count of annotations. (The average duration of annotating a document was 57 seconds in the final annotation round.)

As can be seen in Table 2, the minimal acceptable value of Krippendorff's alpha was not reached. Getting a high agreement in web genre classification is hardly possible. Defining genres both interesting for web corpus users and agreeable to annotators is difficult. That is our conclusion as well as others: [3, 22].

If there is a reasonable solution, it must require a different approach or a lower level of target inter-annotator agreement or both. We suggest taking these measures to get a reliable genre annotation of web corpora:

- Remove paragraphs showing sings of a genre different from the major genre in the training data (such paragraphs were marked by annotators).
- Continue the process with active learning rounds to efficiently annotate more data.

- Consider using whole single genre web sites for training. This technique helped in our other work on non-text removal and topic classification of an Estonian web corpus.
- Train the classifier only on documents with a perfect agreement.
- Set a high top label probability threshold of the classifier. That will increase precision at the cost of recall. The users of text corpora will not mind if the genre of some documents remains unknown. They mind the precision.

Furthermore, since the borders of genres are not strict, we suggest a different approach to the evaluation: the 'User's point of view'. To evaluate the classification, corpus users would be asked to assess the genre annotation of random web pages from the corpus (in the plaintext format) by assigning one of the following three labels to each selected document:

1. This is the genre.
2. This could possibly be the genre.
3. This could not be the genre.

We consider 5 % of texts marked as 'This could not be the genre' an acceptable level of classification mistakes.

6 Conclusion

We presented an inter-annotator agreement driven annotation setup for supervised learning of genres in an English web corpus.

To conclude, we will continue the research on genre classification of large web corpora despite the issues described in this paper since we are interested in exploring the genre composition of the English web and offering the corpus users a possibility to focus their research on particular genres.

A keyword comparison of single genre subcorpora to the whole corpus proposed by [10] could also show interesting properties of web genres.

Exploring possibilities for transfer of the method to other languages than English is another topic for further research.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101 and by the Grant Agency of CR within the project 18-23891S.

This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

Appendix: Annotation Manual and Recognised Genres

Definition of genre and classes to recognise from the annotation manual used for the reference of annotators in the '9 genres' annotation scheme follow. This manual is an extension of Sharoff's description of Functional Text Dimensions in [21, pp. 94–95]. Classes were re-organised in a 9 class scheme, linguistic markers to observe in the text were provided and examples of text were added by the author of this paper.

Definitions of Genre for Annotators

- *A particular style or category of works of art; esp. a type of literary work characterised by a particular form, style, or purpose.* – A general OED definition.
- *A set of conventions (regularities) that transcend individual texts, helping humans to identify the communicative purpose and the context underlying a document.* – Santini, Mehler, Sharoff: Genres on the Web: Computational Models and Empirical Studies. Vol. 42. Springer Science & Business Media, 2010.
- Sketch Engine perspective: The users need to know what types of text are included in the corpus. Since the users do language research, building dictionaries, n-gram models for writing prediction etc., including genre information allows them to use subcorpora limited to a particular genre.

The aim of this annotation project is to identify genres in texts on the web. **Genres are not topics.** Topics are determined by content words. A genre is determined by the **style of writing**, content words are only supporting evidence. Therefore it is the style that is key in assessing the genre, content words are only secondary.

Instructions for Annotators

General instructions:

1. You can start or stop working any time. Use always the same nickname.
2. The total count of documents in the annotation round and your progress can be seen on the top of the page. Proceed until 'No more documents to annotate' is displayed.
3. Make breaks frequently. Trying to stay focused for a long time leads to mistakes.
4. Do not hesitate to ask when in doubt. If you refer to a particular document, include the document ID.

Instructions on determining the genre:

5. Read the whole text in case of short or medium sized documents. Read several paragraphs from the beginning, several paragraphs in the middle and several paragraphs from the end of a long document.

6. Is it a text consisting of whole fluent sentences in English? If not, select **Non-text**.
7. Find the largest part of the text bearing the signs of a genre. Are there multiple genres strongly and equally represented? If yes, select **Multiple genres**.
8. Focus on the style of writing. Determine the main genre: **Information** – **Story telling** – **Instructions** – **News** – **Legal** – **Discussion**.
9. If the genre is **Information**, determine the sub-genre: **Promotion** – **Academic** – **Review**.
10. If you are not sure about a genre, read its definition again.
11. It is ok if you still don't know: Mark **Short text** (i.e. too short to determine a genre) by **Strongly** and continue.

Instructions on marking the strength of a presence of a genre in the text:

12. Mark the genre represented the most in a document.
13. Use **Strongly** for the main genre.
14. Use **Partly**/**Somewhat** for less represented genres. Usually not needed.

Instructions on paragraph removal:[7]

15. Use the **Remove** button next to a paragraph to remove it from the current document.
16. Remove Non-text paragraphs.
17. Remove minority genre paragraphs in case there is more than a small number of them.

Recognised Genres (Based on Sharoff's Functional Text Dimensions)

Information = To what extent does the text **provide information**? Examples: topic **definition** (textbooks, general information, encyclopedia), **blogs** (topic blogs, argumentative/point of view blogs), **research** (scientific papers, popular science), advertisement/**promotion** of goods/services/thoughts.

- Subcategory **Promotion** = To what extent does the text **promote a product**, service, political movement, party, religious faith? Examples: A company landing page, an advertisement, a product presentation page, an e-shop catalogue page, a job offer, a page describing the service of a charity, a religious tract, a political manifesto.
- Subcategory **Academic** = To what extent would you consider the text as representing **research**? Usually formal, first person plural, scientific terms. Example: A research paper or any text written using the **academic style**. Also, it can be Partly if a news text reports scientific contents.

[7] The purpose of the paragraph removal function is to make the future training data less noisy.

- Subcategory **Review** = To what extent does the text **evaluate** a specific entity **by endorsing or criticising** it? Usually a personal experience of the reviewer, comparison to other products, pros and cons. Example: product review endorsing or criticising the product.
- Other informative text – select the general **Information**

Story telling = To what extent is the text's content **fictional** or **telling a story**? Examples: **Description of events** in the order they followed (real or fictional), sometimes **informal style**, can be in the first person. All narrative texts belong to this category: fiction, narrative blogs. Example: "I visited New York, I was at the White House, saw the Statue of Liberty, my luggage got lost on the way back."

Instructions = **Teaching the reader** how something works. The **imperative** is frequent. Example: "Fill in all fields in this form and click the OK button. Then wait for three minutes and add one teaspoon of sugar."

News = **Informative report** of events recent (or coming in the near future) at the time of writing (not a discussion or a general state of affairs). Frequently **formal style**, set to a particular place and time. Often quotes sources of information. A diary-like blog entry is also considered **reporting**. Examples: "Prague, 10/28/18. President Zeman said ¡quotation¿." Or: "'Almost five million tourists visited Czech Republic the last year', said Jana Bartošová, the deputy of the minister of culture."

Legal = To what extent does the text lay down a **contract** or specify a **set of regulations**? Examples: a law, a contract, copyright notices, university regulations.

Discussion = A written communication of participants of a discussion. Frequently **personal** and **informal style**. Examples: expressing points of view, giving advice, responses/comments to the original article or previous comments, sharing personal experiences. Can be **multiple authors**. (Note that just describing how something works is Information, just giving instructions how to solve a problem is Instructions.)

Non-text = To what extent is the text different from what is expected to be a normal running text? Examples: Lists of links, online forms, tables of items, bibliographic references, cloud of tags, sentences ending in the middle, machine generated text. Not a Non-text if at least 30 words belong to nice whole sentences.

Short text/Unsure = A valid text (not a Non-text) that is too short to determine its genre. Or a text not belonging strongly to any class.

Multiple genres = A valid text (not a Non-text) consisting of several long parts showing strong signs of multiple genres. Example: A long news article with a long discussion below the article. Instruction: Select Multiple genres, then mark particular genres by Partly. Use the Remove button to remove paragraphs of a minor genre instead.

References

[1] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

[2] Douglas Biber. "A typology of English texts". In: *Linguistics* 27.1 (1989), pp. 3–44.

[3] Kevin Crowston, Barbara Kwasnik, and Joseph Rubleske. "Problems in the use-centered development of a taxonomy of web genres". In: *Genres on the Web*. Springer, 2010, pp. 69–84.

[4] Erika Dalan and Serge Sharoff. "Genre classification for a corpus of academic webpages". In: *Proceedings of the 10th Web as Corpus Workshop*. 2016, pp. 90–98.

[5] Johan Dewe, Jussi Karlgren, and Ivan Bretan. "Assembling a balanced corpus from the internet". In: *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*. 1998, pp. 100–108.

[6] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. "Introducing and evaluating ukWaC, a very large web-derived corpus of English". In: *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*. 2008, pp. 47–54.

[7] Richard Forsyth and Serge Sharoff. "Document Dissimilarity within and across Languages: a Benchmarking Study". In: *Literary and Linguistic Computing* 29 (2014), pp. 6–22.

[8] Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, Vít Suchomel, Jan Bušta, Vít Baisa, and Jan Michelfeit. "The TenTen Corpus Family". In: *7th International Corpus Linguistics Conference CL*. UCREL. Lancaster, 2013, pp. 125–127.

[9] Jelena Kallas and Kristina Koppel. *Eesti keele ühendkorpus 2019*. Centre of Estonian Language Resources.

[10] Adam Kilgarriff. "Getting to know your corpus". In: *International conference on text, speech and dialogue*. Springer. 2012, pp. 3–15.

[11] Adam Kilgarriff and Tiberius Carole. *Genre in a frequency dictionary*. Presented at the International Conference on Corpus Linguistics, Lancaster, 2013. 2013.

[12] Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, et al. "SYN2015: Representative corpus of contemporary written Czech". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2522–2528.

[13] Klaus Krippendorff. *Content analysis: An introduction to its methodology (2nd ed.)* Thousand Oaks, CA: Sage, 2004.

[14] David Y. W. Lee. “Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle”. In: *Language Learning and Technology* 5.3 (2001), pp. 37–72.

[15] Geoffrey Leech. “100 million words of English: the British National Corpus (BNC)”. In: *Language Research* 28.1 (1992), pp. 1–13.

[16] David D Lewis and Jason Catlett. “Heterogeneous uncertainty sampling for supervised learning”. In: *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.

[17] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[18] Alexander Mehler, Serge Sharoff, and Marina Santini. *Genres on the web: Computational models and empirical studies*. Vol. 42. Springer, 2010.

[19] Mark Rosso. “User-based identification of Web genres”. In: *Journal of the American Society for Information Science and Technology* 59.7 (2008), pp. 1053–1072.

[20] Burr Settles. *Active Learning*. Vol. 6. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

[21] Serge Sharoff. “Functional text dimensions for the annotation of web corpora”. In: *Corpora* 13.1 (2018), pp. 65–95.

[22] Sven Meyer Zu Eissen and Benno Stein. “Genre classification of web pages”. In: *Annual Conference on Artificial Intelligence*. Springer. 2004, pp. 256–269.