

Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing

Noushin Rezapour Asheghi, Serge Sharoff, Katja Markert

School of Computing, School of Modern Languages and Cultures, School of Computing
University of Leeds United Kingdom
scs5nra@leeds.ac.uk, s.sharoff@leeds.ac.uk, markert@leeds.ac.uk

Abstract

Research in Natural Language Processing often relies on a large collection of manually annotated documents. However, currently there is no reliable genre-annotated corpus of web pages to be employed in Automatic Genre Identification (AGI). In AGI, documents are classified based on their genres rather than their topics or subjects. The major shortcoming of available web genre collections is their relatively low inter-coder agreement. Reliability of annotated data is an essential factor for reliability of the research result. In this paper, we present the first web genre corpus which is reliably annotated. We developed precise and consistent annotation guidelines which consist of well-defined and well-recognized categories. For annotating the corpus, we used crowd-sourcing which is a novel approach in genre annotation. We computed the overall as well as the individual categories' chance-corrected inter-annotator agreement. The results show that the corpus has been annotated reliably.

Keywords: genres , annotation, crowd-sourcing

1. Introduction

In approaching a collection of texts, it is very natural to ask the question: what kinds of texts does it contain? Attempts to categorise texts by their genre go back to Aristotle (Santini et al., 2010). Detecting the genre of a text is beneficial in many areas of Natural Language Processing. For example, in POS tagging, machine translation or discourse annotation knowing the genre of a document can help in selecting more appropriate language models. Giesbrecht and Evert (2009) showed that change in the genre of a dataset can have a direct impact on accuracy of POS tagging. In their experiment, the POS tagging achieves 96.9% accuracy on newspaper texts whereas it reaches only 85.7% accuracy on forums. Webber (2009) showed that genres such as letters to the editor vs. newspaper articles differ in the distribution of particular discourse relations. In Information Retrieval users often find it difficult to find relevant pages that are in the right genre (Vidulin et al., 2007). Therefore automatic genre identification can be employed in search engines in order to improve the search results.

The interest in the Web and its genres (Mehler et al., 2010) resulted in a proliferation of genre-annotated corpora, each of which was built according to its specific principles, using its own classification scheme and annotation guidelines. In this paper we will review previous attempts at collecting these corpora, assess their shortcomings and limitations especially the problem with their reliability and present a new annotated corpus created with the aim of achieving high inter-annotator agreement on an arbitrary web page.

2. Existing Web Genre-annotated Corpora

Several efforts have been made to build genre annotated web corpora and to employ them for research in the field of automatic genre identification (AGI). But each collection is different in terms of the size of the corpus; how the web pages were collected; how the web pages were preserved

and the set of genre labels used. Table 2 compares some properties of these corpora. The following is a summary of each genre collection.

Hierarchical Genre Collection (HGC) (Stubbe and Ringlsetter, 2007) was annotated based on a set of hierarchical genre labels with seven main categories and thirty two sub categories, e.g., *literature* as a main category with the subcategories *poem*, *prose* and *drama*. This collection consists of 1280 web pages preserved in HTML format. For each genre category, forty example pages were manually collected.

I-EN-Sample (Sharoff, 2010) consists of 250 web pages randomly selected from I-EN corpus of 71,636 pages representing a snapshot of the English Web (Sharoff, 2006). The collection was annotated using the Functional Genre Classification (FGC) scheme which consists of seven macro-genres aimed at describing any text. The genre palette in FGC is based on the function or the purpose of the document, e.g., *instruction* which covers *FAQs*, *manuals* and *tutorials*.

KI-04 (Meyer zu Eissen and Stein, 2004) is another genre-annotated web corpus consisting of 1209 HTML documents. This collection has been annotated using eight genres, e.g., *link collection*, *shop* and *articles*. The genre list in this collection was developed based on the result from a study of usefulness of genre classes, which was determined by asking a group of students to fill a questionnaire about the typical topics for queries and favourite genre classes.

The KRYS I (Berninger et al., 2008) collection consists of 6200 PDF documents. This corpus has been annotated using seventy genres which are grouped into ten sets, e.g. *commentary* and *review* in the *journalism* group. Although this selection is meant to be a web genre-annotated corpus, it includes only web pages in the PDF format. Therefore, genres that do not normally use this format, such as *homepage* and *shop*, are not included in this corpus.

MGC (Multi-labelled Genre Collection) (Vidulin et al.,

Annotators	Agreement
Student and Secretary I	51.74%
Student and Secretary II	53.76%
Secretary I and II	45.65%
All three	37.85%

Table 1: Human agreement for the KRYIS I corpus (Berninger et al., 2008) which has *seventy* genre classes. Results illustrate a low percentage agreement.

2007) is the only genre-annotated corpus which allowed multi-labelling. This means that each web page can be categorized as belonging to several genre classes. It consists of 1539 web pages classified into twenty genres. They were collected by targeting web pages in these genres, as well as using random web pages and popular web pages coming from Google Zeitgeist.

SANTINIS (Santini et al., 2007) corpus which consists of 1400 web pages was annotated based on seven genres. This collection focused on genres which are exclusive to the web, e.g. *blogs* and *FAQs*. In the compilation of this corpus only web pages which clearly belong to these genres were manually collected.

The Syracuse (Crowston et al., 2011) collection consists of 3027 web pages annotated based on 292 genres. The genre palette in this collection was developed by asking three groups of people (teachers, journalists, engineers) to produce web genre terms themselves.

Reliability One problem with all the existing genre-labelled collections is the issue of reliability of the annotations. Corpora such as SANTINIS, KI-O4 and Syracuse have been annotated by a single person and as a result, their inter-annotator agreement measures cannot be computed. The MGC, I-EN-Sample and KRYIS I corpora have been double-annotated. However, agreement measures were low ($\alpha=0.56$ for MGC and $\alpha=0.55$ for I-EN-Sample) for the part of the corpora which have been selected randomly from the web (Sharoff et al., 2010). Even for the KRYIS-I corpus, which has not been selected randomly, Table 1 shows a low percentage agreement, and potentially lower chance-corrected measures (Artstein and Poesio, 2008).

Size Moreover, these collections are not large enough to ensure representativeness of genre classes. Table 2 compares these collections in terms of maximum, minimum and median number of web pages per genre category. They often have few annotated web pages per category, especially for the KRYIS-I and Syracuse collections, while machine learning algorithms often require a reasonable number of training examples in order to produce satisfactory results.

Format Another major drawback of some of the existing corpora is that they have been preserved in different formats such as PDF or plain text which results in losing HTML tags. For instance, each web page in KRYIS I corpus is saved in PDF format and as a result automated tools are needed to convert PDF to plain text or HTML format. However, these tools are error prone and as a result some information may be lost or wrongly converted. Also previous studies in AGI show that HTML tags can improve the

accuracy of genre classification (Kanaris and Stamatatos, 2009).

Topic Diversity Also, some of these corpora have been collected from a small number of sources which are topically similar. For example web pages in the genre class *frequently asked questions* in Santinis corpus (Santini et al., 2007) are mostly about *hurricane*. However, a corpus without any false correlation between genres and topics is needed in order to develop a learning model which can detect the genres of web pages without being influenced by the spurious connections between genres and topics.

3. Building a Reliable Genre-annotated Corpus

Currently there is no reliable genre-annotated corpus of Web pages with a large number of manually annotated documents. The drawbacks of existing genre-annotated web corpora (i.e., relatively small number of web pages per genre category; low inter-coder agreement; pages collected from small number of sources; preserved in different formats such as PDF or plain text) highlight the necessity of developing a reliable genre-annotated corpus. Therefore, we have designed and built a web corpus which fulfils the following criteria:

- It needs to be reliable (chance corrected agreement rate for this corpus needs to be high).
- It must be collected from a diverse range of sources in order to avoid creating false correlations between genres and topics.
- It must include genre classes which are exclusive to the web.
- Web pages must be saved in HTML format. Also the appearance of each web page must be preserved by taking a screen shot of its whole content.

3.1. Definition of Genre Classes

The quality of manual annotations depends on the use of precise and consistent guidelines which include the definitions of the categories. Therefore, the development of the annotation guidelines must be seen as one of the crucial tasks in annotation projects. The vagueness and ambiguity in the annotation guidelines especially the definition of the categories which increases the subjectivity of the annotation task could be the reason for low inter-coder agreement in existing web genre corpora. For example, annotators may have different interpretation of broad and vague categories such as *informative* and *entertainment* in MGC corpus (Vidulin et al., 2007).

Therefore, to ensure the success of our annotation task we defined the categories in the annotation guidelines in a way that enables humans to adequately differentiate among them and avoided ambiguous and vague categories. Since we intended to build a web genre corpus, we gave priority to genres which are exclusive to the web such as *home pages*, *frequently asked questions* and *personal blogs*. We

Corpus	Number of		Number of pages per genre			Format	Collection method	Reliability
	pages	genres	min	max	median			
KRYS I (Berninger et al., 2008)	6200	70	6	117	97	PDF	focused search	a.p.a.= 50.38% (Table 1)
MGC (Vidulin et al., 2007)	1536	20	55	227	77	HTML with images	both random selection and focused search	Low $\alpha=0.56$ for the random web pages(Sharoff et al., 2010)
HGC (Stubbe and Ringstetter, 2007)	1412	32	40	40	40	HTML only	focused search	not measured
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8	126	205	145	HTML only	focused search	not measured
SANTINIS (Santini et al., 2007)	1400	7	200	200	200	HTML only	focused search	not measured
I-EN-Sample (Sharoff, 2010)	250	7	10	99	30	TXT from HTML	random selection	Low $\alpha=0.55$ (Sharoff et al., 2010)
Syracuse (Crowston et al., 2011)	3027	292	1	174	3	HTML only	focused search	not measured

Table 2: Summary of genre-annotated corpora. a.p.a. stands for average percentage agreement.

also focused on genres which most web genre corpora include them such as *news*. Moreover, we included categories such as *editorial* and *review* in order to test the capability of classifiers in distinguishing facts from opinions. Table 3 shows the set of 15 genre labels and their definitions used in our genre annotation task, while, Table 4 shows how these 15 selected genre classes correlate with those used in other genre-annotated corpora. However, since, different genre-annotated corpora used different genre classes with different level of granularity, the one-to-one comparison between our genre labels and their genre classes might not be feasible. For example, the genre label *journalistic* in MGC can include several genre in our corpus such as *news*, *editorial*, *interviews* and *reviews*. Another example could be *periodicals* (newspaper, magazine) from KRYS I corpus which is very broad and can include many genre classes such as *recipe*, *interview* and *reviews*.

3.2. Corpus Compilation

The next step after defining the categories is corpus compilation. Web corpora are categorized into two subtypes, i.e. designed and crawled, which are different in terms of how they are collected (Kilgarriff, 2012). The content of a designed corpus is selected based on its design specification whereas there is much less control on the content of a corpus constructed by crawling the web. HGC (Stubbe and Ringstetter, 2007) and UKWac (Baroni et al., 2009) are examples of designed and crawled corpora respectively. We chose to build a designed corpus for two reasons. First, we wanted a balanced corpus with a large number of web pages for each category. While crawling the web is a cheap and fast way of collecting web pages, there is no guarantee that it fulfils this criterion. Second, crawled web pages could be noisy whereas, manually collected clear and prototypical examples are better for machine learning. Use of a designed corpus was also suggested by Rehm et al. (2008) as an initial step in building a reference corpus of web genres. However, one disadvantage of designed corpus is that it could overestimate the agreement. In other words, reaching a high inter-coder agreement for genre annotation could be more difficult in random web pages.

Therefore, in order to obtain a balanced collection, we

hand-selected web pages mainly from Yahoo Directory¹ and Open Directory Project² websites. We tried to select web pages from a diverse range of sources to avoid creating false correlation between topic and genre labels (see source diversity of the corpus summarized in Table 9).

In the next phase, we saved the pages in HTML format using KrdWrd (Steger and Stemle, 2009). However, only saving a web page in HTML format does not guarantee preservation the appearance of a web page. To achieve this, we can either save the graphic and style files of each page, or take its screen shot. We chose the second option and used KrdWrd to preserve each web page as an image.

3.3. Annotation Procedure

After the compilation of the web pages, the corpus needs to be annotated with the set of chosen genre labels which can be a very time consuming and expensive task. However, in recent years, the advent of

-sourcing (e.g. via Amazon Mechanical Turk³) has facilitated annotation tasks and this phase can be done cheaper and faster than ever before. Amazon Mechanical Turk (MTurk) has been used for a variety of labelling and annotation tasks e.g. word sense disambiguation, word similarity, text alignment, temporal ordering (Snow et al., 2008); machine translation (Callison-Burch, 2009); building question answering dataset (Kaiser et al., 2008); but not for genre annotation.

3.3.1. Amazon's Mechanical Turk

The Mechanical Turk web site provides a service which enables requesters such as researchers or companies to create and publish jobs also known as Human Intelligence Tasks (HITs). These HITs can be done by untrained MTurk workers (turkers) all around the world for a small amount of money. The main advantages of Mturk are low cost and efficiency in terms of the speed of task completion as well as its infrastructure which allows the requesters to develop their HITs using standard HTML and Javascript.

As turkers are motivated by profit, quality control of the result is crucial in order to detect poor quality or randomly selected answers. Moreover, Mturk HITs like any other

¹<http://dir.yahoo.com/>

²<http://www.dmoz.org/>

³<https://www.mturk.com/mturk/welcome>

Genre	Definition
Personal Homepage (php):	created by an individual to contain content of a personal nature rather than on behalf of a company, organization or institution.
Company/ Business Homepage (com):	the main web page of a company or an enterprise website which promote a product or a service. These web pages often contain a description of the purpose or objectives of the company.
Educational Organization Homepage (edu):	the main web page of an educational institution website. Examples are universities and schools home pages.
Personal Blog /Diary (blog):	where people write about their day-to-day experiences (please only choose this option if the blog is personal and it is about personal experiences)
Online Shops (shop):	Web pages created with intention to sell
Instruction/ How to (instruction):	contains instructions and teaches you how to do something (not recipes)
Recipe:	a set of instructions that describe how to prepare or make food
News Article (news):	a report of recent events
Editorial:	an opinion piece written by the editorial staff or publisher of a newspaper or magazine
Conversational Forum (forum):	where people have a conversation about a certain topic
Biography (bio):	a detailed description of someone's life
Frequently Asked Questions (faq):	listed questions commonly asked about a particular topic
Review:	an evaluation of a publication, a product or a service, such as a movie,a video game, a musical composition or a book
Interview	a conversation in which one or more persons question another person
Story	a narrative, either true or fictitious, with the aim to entertain the reader

Table 3: Definition of genre labels. To save the space, in this paper we use the abbreviation of genre labels which are specified in front of the genre names.

web-based interface are vulnerable to automated scripts also known as bots which are used by some workers in order to maximize their income from Mturk (Mason and Suri, 2012). To ensure a high quality result, Mturk provides two types of qualification criteria which a requester can add to the HIT design. The first type which is referred to as “system qualifications” includes HIT submission rate (the percentage of submitted HITs by the worker), HIT approval rate (ratio of accepted HITs compared to the total number of HITs submitted by the worker), HIT rejection rate (ratio of rejected HITs compared to the total number of HITs submitted by the worker) and location (the worker’s country of residence).

The second type of quality control measures is a qualification test which can be designed by the requesters based on their tasks as well as the skills and the knowledge they are seeking in the workers. Up to five qualification criteria can be assigned to a HIT by the requester and therefore, only workers who pass these qualification measures are permitted to complete the HITs. In addition, Mturk enables the requesters to download and review the submitted works and then reject poor quality data and only pay for the HITs which they approve. In the next section which describes HIT design, we use both “system qualifications” and “qualification test” in order to ensure the quality of the annotations.

3.3.2. HIT Design

This section describes the details of HIT design and quality control measures which were developed to ensure obtaining better quality data. In order to keep the annotation task simple, we decided to choose the single-labelling method, despite the fact that, there are some web pages that belong to

Genre	KRYS I	MGC	HGC	KI-04	SANTINIS	Syracuse
php		✓	✓	✓	✓	✓
com		✓				✓
edu						
blog		✓	✓	✓	✓	✓
shop		✓	✓	✓	✓	✓
instruction				✓		✓
recipe						✓
news	✓		✓			✓
editorial			✓			✓
forum	✓	✓	✓	✓		✓
bio	✓		✓			✓
faq	✓	✓	✓		✓	✓
review	✓		✓			✓
interview	✓		✓			✓
story	✓	✓	✓			✓

Table 4: This table illustrates which genre classes in our corpus are also included in existing genre-annotated corpora.

more than one genre class (Crowston and Kwasnik, 2004; Kessler et al., 1997; Santini, 2008). Therefore, one of the defined genre labels in the guidelines or the option “other” can be chosen for each web page.

In the HIT implementation phase, the simplest approach is to implement one annotation task (one web page with the choice of all the genre labels) per HIT. However, we decided to design the HITs in a way that each HIT includes annotating ten web pages. This enables us to use one of ten annotation tasks in a HIT as a quality control “trap” question for identifying the workers who select the answers randomly. A set of twenty web pages from dif-

ferent genre classes in our dataset which the first author of this paper judged them as unambiguous and clear example of one of our predefined genre categories, were selected as gold standard and used as trap questions. We performed semi-automated monitoring of the annotations by checking the answers to the trap questions and rejected the work from workers who gave wrong answers to the trap questions more than 80% of the time.

Also, to ensure reliable and high quality data, we restricted the range of workers who can complete our task. We only allowed workers who had completed at least fifty previously accepted HITs; have approval rate higher than 95% and pass our qualification test with the score of equal or higher than 80%. The qualification test includes the definitions and examples of genre classes as well as ten multiple-choice genre annotation questions.

Because adding more annotators can help to reduce annotation bias, it is encouraged in human annotation projects to have as many annotators as possible (Beigman Klebanov and Beigman, 2009). We chose to have five annotations per web page, because, Snow et al. (2008) compared the quality of annotation done by experts and Mturk workers and concluded that an average of 4 non-expert workers in Mturk often provides expert-level label quality.

3.4. Inter-coder Agreement Measures

In Natural Language Processing and machine learning, a reliably annotated dataset plays a crucial role. Reliability of annotated data is an essential factor for reliability of the research result. In other words, the results of research based on unreliable annotation can be considered as untrustworthy, doubtful and even meaningless. In order to be able to measure the reliability of annotation, different annotators judge the same data and the inter-coder agreement is calculated for their judgements. The most commonly used inter-coder agreement measure which employed to measure the extent of consensus in judgements among annotators are: Percentage agreement, S (Bennett et al., 1954), Scott's π (Scott, 1955), Cohen's κ (Cohen and others, 1960) and Krippendorff's α (Krippendorff, 1970).

Percentage or observed agreement which is the simplest measure of agreement among coders can be computed by simply summing the number of instances on which the annotators agree and dividing it by the total number of instances.

Although the computation of observed agreement is not complicated, this measure cannot be trusted because it does not take into account the agreement which is expected to happen by chance and as a result it can overestimate the true agreement. Therefore, in order to overcome the shortcoming of percentage agreement, other inter-coder agreement measures such as Scott's π or Cohen's κ which correct for chance agreement must be computed. Originally these coefficients were proposed for calculating inter-coder agreement between two annotators. Then Fleiss (Fleiss, 1971) proposed a generalization for Scott's π and Davies and Fleiss (Davies and Fleiss, 1982) gave generalization for Cohen's κ . Although these two measures are very similar and often have very close values, there is one difference between them. For calculating expected agreement for π

Genre Labels	Percentage agreement	π
Personal Homepage	0.979	0.858
Company/ Business Homepage	0.962	0.713
Educational Organization Homepage	0.993	0.953
Personal Blog /Diary	0.977	0.812
Online Shops	0.976	0.830
Instruction/ How to	0.985	0.871
Recipe	0.995	0.971
News Article	0.970	0.801
Editorial	0.981	0.877
Conversational Forum	0.994	0.951
Biography	0.988	0.905
Frequently Asked Questions	0.992	0.915
Review	0.984	0.880
Story	0.996	0.953
Interview	0.992	0.905

Table 5: Inter-coder agreements for individual categories show substantial agreement among the coders. Therefore annotations for all the genre classes are highly reliable.

we only take into account the combined judgements of all coders and not the number of items assigned to each category by each individual coder. Unlike π , for calculating expected agreement for κ , we take into account the number of times each coder assigns an item to a category.

Since in Mturk the annotations have been done by various workers, κ is not a good measure as it takes into account the proportion of items assigned by each annotator to each category. Therefore, like other annotation studies using crowdsourcing (e.g. (Mohammad and Turney, 2012; McCreddie et al., 2011; Bentivogli et al., 2011)) we calculated a generalization of π also known as Fleiss's kappa (Fleiss, 1971) for the annotation. The next section presents the result of inter-coder agreement results.

3.5. Results of Annotation Study

The annotation task was completed within seven days with the total cost of \$820. Overall 42 annotators participated in annotating the corpus in Mturk. The annotation study shows high agreement in the annotation results. The percentage agreement is 88.2% and π is 0.874. Based on the interpretation of the inter-coder agreement value by Landis and Koch (Landis and Koch, 1977), the π value for our annotation task shows perfect agreement between the annotators and therefore we can consider the annotation reliable. We also computed π for each single category in order to identify the most and the least agreed on categories. Single category π measures the agreement for one target category and treats all other categories as one non-target category and measures agreement between the two resulting categories. Table 5 shows the results of inter-coder agreement measures for individual genre classes. Results show that *recipe* was the easiest category for the annotators whereas *company/ business home pages* caused the most disagreement between the annotators. However, π values for the in-

Types of inter-annotator agreement	# of web pages	% of web pages
5,0	2945	74.29%
4,1	791	19.95%
3,1,1	104	2.62%
3,2	116	2.92%
2,1,1,1	4	0.10%
2,2,1	4	0.10%
1,1,1,1,1	0	0%

Table 7: Distribution of different types of inter-annotator agreement

dividual categories illustrate substantial agreement among the coders and, as a result, annotations for all the genre classes are highly reliable. Overall we show that genre identification for the listed genre classes can be reliably annotated and therefore is a well-defined tasks for automatic classification.

The next phase of building a reliable genre annotated dataset for developing supervised machine learning classifiers is to convert the annotated dataset into a gold standard. There are a number of different methods to derive a gold standard from an annotated dataset (Beigman Klebanov and Beigman, 2009). For instance, the annotators can discuss together (Litman et al., 2006) to reach an agreement on the disagreed items or if more than two annotators are employed in the annotation task, a majority vote approach (Vieira and Poesio, 2000) can be employed on the disagreed items. Also, a domain expert can be used to decide the final label for the disagreed instances (Girju et al., 2006; Snyder and Palmer, 2004) or simply the instances which cause disagreement can be excluded from the dataset (Beigman Klebanov and Beigman, 2009).

Since, we employed Mturk for annotation, reaching agreement through discussion between annotators is not possible. Therefore, as we have five annotations per web page, the majority vote strategy was employed to assign the final label to the disagreed web pages. There are seven possible types of inter-annotator agreement when there are five annotators (Table 6).

In order to analyse how often the annotators agreed with each other, we calculated the percentage of each type of inter-annotator agreement (Table 7). For more than 74% of the web pages all the five annotators agreed and for 95% of the data at least four annotators agreed which indicates high level of agreement between the coders. Low percentage of the other five types of inter-coder agreement confirms the high value of π for the annotation task. Disagreements in cases where only three annotators agreed with each other are mainly caused by confusion between *news* and *editorial* and between *shop* and *company home page*. Since we did not have majority vote for eight web pages, the final label for these instances were assigned by one of the authors.

Number of genres	15
Number of web pages	3964
Number of web pages for the smallest category	184
Number of web pages for the largest category	332
Median Number of web pages for the categories	266
Number of tokens	7,205,820
Number of types	130,254
Number of sentences	329,861

Table 8: The corpus statistics

4. Corpus Statistics

In order to provide further insight into the constructed corpus, we computed some corpus statistics such as number of tokens, number of types and number of sentences. Table 8 gives an overview of the corpus statistics, while Table 9 shows the source diversity of the corpus. The corpus consists of 3964 web pages, distributed across 15 genres. It contains more than 7 million words which makes it approximately seven times bigger than the Brown corpus in terms of the number of tokens. Also, in order to investigate which dates these web pages were published or last modified, I used Stanford Named Entity Recognizer (Finkel et al., 2005) to identify all the dates in each page. Then the latest date was taken as the publish date or last modified date. The results show that about 75% of the web pages were last updated in years 2010 to 2012.

Genre	Number of		Number of pages from the same website		
	web pages	websites	max	min	med
php	304	288	9	1	1
com	264	264	1	1	1
edu	299	299	1	1	1
blog	244	215	9	1	1
shop	292	209	23	1	1
instruction	231	142	15	1	1
recipe	332	116	8	1	1
news	330	127	12	1	1
editorial	310	69	11	1	3
forum	280	106	11	1	1
bio	242	190	15	1	1
faq	201	140	8	1	1
review	266	179	15	1	1
story	184	24	38	1	7
interview	185	154	11	1	1

Table 9: Statistics for individual categories which illustrate source diversity of the corpus. Max, min and med are abbreviations of minimum, maximum and median respectively.

5. Conclusions

To the best of our knowledge, we present the first web genre corpus which is reliably annotated.⁴ We developed precise

⁴A plan for developing another reliable genre corpus of 50,000 web pages has been recently announced (Egbert and Biber, 2013), but no results have been reported so far.

Types of inter-annotator agreement	Represented as
all agreed on a choice of category	5,0
four annotators agreed and the fifth disagreed	4,1
only three annotators agreed with each other while the other two disagreed with the majority as well as each other	3,1,1
only three annotators agreed with each other while the other two disagreed with the majority but agreed with each other	3,2
only two annotators agreed with each other	2,1,1,1
two annotators chose the same category and the other two annotators also chose the same category but different from the first two annotators	2,2,1
all five annotations differed	1,1,1,1,1

Table 6: All possible combination of five annotations

and consistent annotation guidelines which consist of well-defined and well-recognized categories. For annotating the corpus, we used crowd sourcing which is a novel approach in genre annotation. The result of inter-coder agreement shows that the corpus has been annotated reliably. The future work involves extending this corpus by using random web pages. We also plan to extend the number of genre classes. Researchers in genre classification have come up with long lists of genre classes, e.g., 292 genre labels in the Syracuse corpus (Crowston et al., 2011) or 500 genre labels (Dimter, 1981). Nevertheless, the list of genre categories is open-set and is never going to be complete since the new ones are emerging all the time.

6. Acknowledgements

This research was partly funded by the Google Research Award to Serge Sharoff and Katja Markert, as well as by EU FP7 funding, contract No 251534 (HyghTra). Also Noushin Rezapour is funded by an EPSRC Doctoral Training Grant.

References

- R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- B. Beigman Klebanov and E. Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.
- E.M. Bennett, R. Alpert, and AC Goldstein. 1954. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303.
- L. Bentivogli, M. Federico, G. Moretti, and M. Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. *Proceedings of the MT Summit*, 13:521–528.
- V. Berninger, Y. Kim, and S. Ross. 2008. Building a document genre corpus: a profile of the KRY5 I corpus.
- C. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon’s mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.
- J. Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Kevin Crowston and Barbara H Kwasnik. 2004. A framework for creating a faceted classification for genres: Addressing issues of multidimensionality. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 9–pp. IEEE.
- K. Crowston, B. Kwaśnik, and J. Rubleske. 2011. Problems in the use-centered development of a taxonomy of web genres. *Genres on the Web*, pages 69–84.
- M. Davies and J.L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.
- Matthias Dimter. 1981. *Textklassenkonzepte heutiger Alltagssprache: Kommunikationssituation, Textfunktion und Textinhalt als Kategorien alltagssprachlicher Textklassifikation*, volume 32. Walter de Gruyter.
- Jesse Egbert and Douglas Biber. 2013. Developing a user-based method of register classification. In *Proc. 8th Web as Corpus Workshop*, Lancaster, July.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.
- J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.
- E. Giesbrecht and S. Evert. 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In *Web as Corpus Workshop (WAC5)*, page 27.

- R. Girju, A. Badulescu, and D. Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.
- M. Kaisser, M. Hearst, and J.B. Lowe. 2008. Evidence for varying search results summary lengths. In *Proc. of ACL*.
- I. Kanaris and E. Stamatatos. 2009. Learning to recognize webpage genres. *Information Processing & Management*, 45(5):499–512.
- B. Kessler, G. Numberg, and H. Schutze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics.
- A. Kilgarriff. 2012. Getting to know your corpus. In *Text, Speech and Dialogue*, pages 3–15. Springer.
- K. Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61.
- J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.
- D. Litman, J. Hirschberg, and M. Swerts. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Computational linguistics*, 32(3):417–438.
- W. Mason and S. Suri. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods*, 44(1):1–23.
- R. McCreddie, C. Macdonald, and I. Ounis. 2011. Crowdsourcing blog track top news judgments at trec. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 23–26.
- Alexander Mehler, Serge Sharoff, and Marina Santini, editors. 2010. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.
- S. Meyer zu Eissen and B. Stein. 2004. Genre classification of web pages. *KI 2004: Advances in Artificial Intelligence*, pages 256–269.
- S.M. Mohammad and P.D. Turney. 2012. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.
- G. Rehm, M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proc. of the 6th Language Resources and Evaluation Conf.(LREC 2008), Marrakech, Morocco, May*.
- M. Santini, R. Evans, R. Power, and L. Pemberton. 2007. Automatic identification of genre in web pages. *University of Brighton*.
- Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.
- M. Santini. 2008. Zero, single, or multi? genre of web pages through the users’ perspective. *Information Processing & Management*, 44(2):702–737.
- W.A. Scott. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*.
- S. Sharoff, Z. Wu, and K. Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 3063–3070.
- S. Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98.
- Serge Sharoff. 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 149–166. Springer, Berlin/New York.
- R. Snow, B. O’Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.
- B. Snyder and M. Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.
- Johannes M. Steger and Egon W. Stemle. 2009. KrdWrd – architecture for unified processing of web content.
- A. Stubbe and C. Ringlstetter. 2007. Recognizing genres. *Proc. Towards a Reference Corpus of Web Genres*.
- V. Vidulin, M. Luštrek, and M. Gams. 2007. Using genres to improve search engines. *Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, 4:45.
- R. Vieira and M. Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.
- B. Webber. 2009. Genre distinctions for discourse in the penn treebank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 674–682. Association for Computational Linguistics.