Lang Resources & Evaluation (2016) 50:603–641
DOI 10.1007/s10579-015-9331-6

ORIGINAL PAPER

Crowdsourcing for web genre annotation

Noushin Rezapour Asheghi[1] · **Serge Sharoff**[2] · **Katja Markert**[3,4]

Published online: 9 January 2016

Abstract Recently, genre collection and automatic genre identification for the web has attracted much attention. However, currently there is no genre-annotated corpus of web pages where inter-annotator reliability has been established, i.e. the corpora are either not tested for inter-annotator reliability or exhibit low inter-coder agreement. Annotation has also mostly been carried out by a small number of experts, leading to concerns with regard to scalability of these annotation efforts and transferability of the schemes to annotators outside these small expert groups. In this paper, we tackle these problems by using crowd-sourcing for genre annotation, leading to the Leeds Web Genre Corpus—the first web corpus which is, demonstrably reliably annotated for genre and which can be easily and cost-effectively expanded using naive annotators. We also show that the corpus is source and topic diverse.

Keywords Genres on the web · Reliability testing · Annotation guidelines · Crowdsourcing

✉ Noushin Rezapour Asheghi
noushin.rezapour@gmail.com

Serge Sharoff
s.sharoff@leeds.ac.uk

Katja Markert
markert@l3s.de

[1] School of Computing, University of Leeds, Leeds LS2 9JT, UK

[2] School of Modern Languages and Cultures, University of Leeds, Leeds LS2 9JT, UK

[3] L3S Research Center, Leibniz-University Hannover, Hannover, Germany

[4] School of Computing, University of Leeds, Leeds LS2 9JT, UK

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 categorize 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 and discourse annotation, knowing the genre of a document can help in selecting appropriate language models. Thus, Giesbrecht and Evert (2009) showed the impact of genre on POS tagging performance. Their POS tagger 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 discourse relations. Genre detection for web texts can also be helpful in information retrieval: Vidulin et al. (2007) make the point that it is difficult for search engine users to find relevant pages that are in the right genres, when starting from standard topical queries.

Realizing this need for genre annotation, even the Brown Corpus, the very first large computerized corpus created in the 1960s, was based on classification of texts into 15 categories, roughly corresponding to genres, such as Press:Reportage, Press:Editorial, Fiction:Adventure, or Fiction:Love and Romance (Kučera and Francis 1967). The British National Corpus (BNC) contains classification of texts according to a range of genre-related parameters, such as the type of publication (e.g., book or newspaper), audience (specialists or lay persons), as well as an explicit genre classification designed by Lee (2001). With the arrival of the web, it became much easier to collect large corpora. The web also resulted in new genres not available before, such as blogs or Internet shopping sites. However, for many genres which feel unique to the web, there are earlier precursors: for example, one could argue that (personal) blogs have similarities to published personal diaries. Section 2 will review more closely the concept of genre and the relations between web and traditional genres.

The interest in the web and its genres (Mehler et al. 2010) resulted in a proliferation of genre-annotated web corpora, each of which was built according to its specific principles, using its own classification scheme and annotation guidelines. Problematically, these corpora are either not tested for annotation reliability as the focus of work was elsewhere or exhibit low inter-annotator agreement. Rehm et al. (2008) already call for a reliably annotated web genre corpus, preferably using a random snapshot of the web, but do not present an actual corpus. This paper takes steps to remedy this research gap. After reviewing prior web genre corpora in Sect. 3, we summarize their shortcomings: these include reliability problems, provision of few pages for many genre classes as well as the occasional lack in source and topic diversity and appropriate storage formats. We suggest that crowdsourcing is the appropriate method to develop a web genre corpus with high inter-annotator reliability because it allows speedy, accurate and inexpensive genre annotation that detaches the annotation proper from the potential bias of the expert team who developed the guidelines [see also Riezler (2014) for discussing the potential circularity if the same team develops guidelines/terms and annotates].

We then present the Leeds Web Genre Corpus (LWGC) that identifies 15 genre classes reliably via crowd-sourcing. Our genre inventory is detailed in Sect. 5. The LWGC consists of two sub-corpora: The first one (LWGC-B(alanced)) is a designed corpus, where web pages were collected using focused search for specific genres by following links in available web directories before them being submitted to the crowd-sourcing annotation. This method allows us to test our annotation method on a set of web pages with little noise. In addition, it leads to a balanced distribution of genres in the corpus, which is ideal for automatic genre identification via machine learning methods that need sufficient training material for each genre—a property that many existing collections lack. In addition, we collect the corpus from a wide variety of sources, circumventing spurious topic-genre correlations existing in some prior corpora. The LWGC-B(alanced) is described in Sect. 6. Our second corpus (LWGC-R(andom)) then expands our method successfully to a corpus where the pages to be annotated are collected in a more arbitrary way among web pages returned by search engines. The LWGC-R(andom) corpus is described in Sect. 7. This sub-corpus also allowed us to investigate and expand coverage of the underlying genre inventory. However, the emphasis of our paper is not on completeness of the genre inventory but on genre annotation methodology.

Our main contribution is therefore the development of a crowd-sourcing genre annotation method which leads to the first web genre corpus with all of the following properties: demonstrably high inter-annotator agreement, regardless of web page provenance, and achievable by non-expert annotators; a large number of web pages per category; source and topic diversity.

2 The concept of genre

Genre definitions. Many researchers have studied the notion of genre, mostly concentrating on the role that the *form* and the *function* of a document play in defining genre. As an example, Campbell and Jamieson (1978) defined genre as:

> a group of acts unified by a constellation of forms that recurs in each of its members. These forms, in isolation, appear in other discourses. What is distinctive about the acts in genre is a recurrence of the forms together in constellation. (Campbell and Jamieson 1978, p. 20)

In this definition, the emphasis is on a document's *form*. In contrast, Miller (1984, p. 159) argues that the definition of genre must not be limited to the form of the discourse only, but it should also include "the action it is used to accomplish". In other words, texts in a genre class have the same purpose or function as well as similar patterns of form. Biber (1991) also emphasizes the importance of *purpose* in recognizing a genre class by stating:

> I use the term genre to refer to text categorizations made on the basis of external criteria relating to author/speaker purpose. (Biber 1991, p. 68)

Swales (1990)'s definition of genre is in line with Biber's as he also recognizes "purpose" as the principle attribute that instances in a genre class share.

We follow Orlikowski and Yates (1994) who use a more comprehensive definition of genre, which combines function and form:

> a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form. (Orlikowski and Yates 1994, p. 543)

Orlikowski and Yates's (1994) definition also adds a new dimension by clearly stressing that genres must be socially recognizable. In other words, genre classes exist only if they are identifiable by people in society (Andersen 2008).

Web genres and traditional genres. Since this paper focuses on genres on the web, it is important to compare web genres with genres in traditional media. The World Wide Web, which was created in 1989, is a communication medium for retrieving and displaying multimedia hypertext documents (Berners-Lee et al. 1994).

Yates et al. (1997) recognized the advent of a new communications medium as one of the reasons for the emergence of variants of existing genres or of new genres. Shepherd and Watters (1998) introduced the notion of *cybergenre* and proposed a hierarchical taxonomy for classifying the genres on the web compared to traditional genres. According to this classification, cybergenres can be extant (i.e.“ based on existing genres”) or novel (i.e. “not like any existing genre in any other medium”). They give on-line newspapers as an example of extant genres and personal homepages as an example of novel genres.[1] Extant genres are divided into two sub-classes: replicated (i.e.“based on genres existing in other media”) and variant (i.e.“a modification of existing genres”). Novel genres are also separated into two groups: emergent (i.e. “derived but significantly different from existing genres”), and spontaneous (i.e. “never employed in other media”). They refer to personalized newspapers and frequently asked questions as examples of emergent and spontaneous genres, respectively.

Crowston and Williams (2000) proposed a similar categorization for web genres. They conducted a survey on 1000 random web pages and distinguish four different types of genres: reproduced, adapted and novel genres as well as unclassified web pages (see Table 1). Reproduced genres replicate genres in traditional media to a great extent and were found to be the most frequent type (60.6 %). The second type (adapted genres) evolved from existing genres in the paper world by using the capability of the new medium. For example, a list of items which makes use of the hyper-link capability of the web to link to other pages is creating both a *list* and an *index*. As a third type they note novel genres exclusive to the web such as *home pages*. Although the proportion of novel genres in this study is very low, it is possible that nowadays this group of genres comprises a bigger percentage due to additional frequent genres such as microblogs. Pages remained unclassified due to two main reasons: not knowing the name of the genre and the difficulty of determining the purpose of the web page. Some of these unclassified web pages could be examples of genres still in formation. Therefore, in the process of building

[1] Dillon and Gushrowski (2000) also argued that the personal homepage is a novel genre on the web, which has no equivalent in the world of print.

Table 1 Percentage of types of genres found by Crowston and Williams (2000)

Type of genre	Percentage (%)
Reproduced genres	60.6
Adapted genres	28.6
Novel genres	5.3
Unclassified web pages	5.6

a genre-annotated web corpus, we would expect to find some web pages without any genre label.

3 Existing genre-annotated web 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, collection methods and web page storage format. In addition, there is no agreed set of genre labels so that each collections' labels vary according to researchers' priorities and the genre definition chosen (see also Sect. 2). In the following, we give a short description of each genre collection, after which we summarize some characteristics all of them share. Table 2 gives an overview of the properties of these corpora.

The hierarchical genre collection (HGC) (Stubbe and Ringlstetter 2007), the Syracuse corpus (Crowston et al. 2011), KRYS I (Berninger et al. 2008) and the corpus constructed in Egbert and Biber (2013), Egbert et al. (2015) use a relatively large number of genre labels (between 32 and 292 labels), leading to high granularity. Their focus is therefore on high coverage and the construction of a detailed taxonomy. HGC, KRYS I and Egbert et al. (2013, 2015) use a hierarchical structure of genre labels so that also a more coarse-grained classification is available.[2] All of them use labels influenced by both form and function of the document, although some labels used relate only to document function or even to document medium, especially in the coarse-grained classification level. This is especially true for Egbert et al. (2013, 2015). More details on each corpus follow.

The hierarchical genre collection (HGC) (Stubbe and Ringlstetter 2007) was annotated using 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 prototypical pages were manually collected.

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 coarse classes, e.g. *Commentary* and *Review* in the *Journalism* group. Although this

[2] It is worth noting that Egbert et al. (2013, 2015) actually do register classification, which they distinguish from genre classification in being mainly based on function, instead of form. The work is still related enough to include here—in particular their subregisters are almost indistinguishable from genre categories.

selection is meant to be a genre-annotated web corpus, it includes only web pages in PDF format. Therefore, genres that do not normally use this format, such as *homepage* and *shop*, are not included.

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

The corpus constructed in Egbert and Biber (2013) has 1000 random web pages categorized into eight very broad, mainly functionally defined genres or registers (e.g. *description*, *discussion* and *opinion*) and 56 sub-registers (which use both form and function). This corpus was annotated via Amazon Mechanical Turk which is a crowd-sourcing website. Later, this project was extended to 53,000 web pages in Egbert et al. (2015). Therefore, their work is the work most similar to ours with regards to annotation methodology. However, they have a stronger focus on coverage whereas we focus on annotation reproducibility, which is low in their work (see Table 2 and further discussion below on reliability).

Then there is a group of corpora with smaller sets of genre labels, either because the researchers focus less on coverage but more on genres which are of interest to them for a certain application or task (KI-04, SANTINIS) or because the authors attempt to achieve high coverage via a broad set of often purely functional labels without further subdivision (I-EN-SAMPLE, MGC to a degree). We will discuss these next.

KI-04 (Meyer zu Eissen and Stein 2004) and SANTINIS (Santini 2007) are the corpora that are most often used in automatic genre identification work. Their categories are motivated by web search use and web specificity, respectively. KI-04 (Meyer zu Eissen and Stein 2004) contains 1205 HTML documents annotated using eight genres, e.g., *link collection, shop* and *articles*. The genre list in this collection focuses on including genre classes that are most useful for web search—it was developed by asking a group of students to fill in a questionnaire about typical topics for queries and favourite genre classes. As can be seen, the resulting classes are of quite differing granularity. SANTINIS (Santini 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. *blog* and *FAQs*.[3] In the compilation of this corpus only web pages which clearly belong to these genres were manually collected.

The MGC corpus (Vidulin et al. 2007) is the only genre-annotated corpus which allowed multi-labeling, i.e. a page can be categorized into several genre classes. It consists of 1536 web pages classified into twenty genres. Some of these genres are defined on purely functional criteria such as *commercial/promotional* whereas some are using both form and function (e.g. *FAQ*). The corpus was collected by targeting web pages in these genres, as well as using random web pages and popular web pages coming from Google Zeitgeist.

I-EN-Sample (Sharoff 2010) consists of 250 web pages randomly selected from the I-EN corpus of web pages representing a snapshot of the English Web texts from

[3] Some of these genres might have precursors in non-web genres as discussed in Sect. 2.

2005 (Sharoff 2006). It was annotated using the Functional Genre Classification (FGC) scheme which consists of seven macro-genres aimed at describing the genre of any text. The genre palette in FGC is based solely on the function or purpose of the document e.g., *discussion* which includes *academic papers, forums, emails* or *political debates*, or *instruction* which covers *FAQs*, *manuals, tutorials*. Therefore, this annotation scheme differs from others by sacrificing depth and specificity of the annotation scheme for coverage.

We are now going to discuss areas of research where we think that the current corpora, regardless of all their diversity, leave open questions and where we can address the corresponding research gap.

Reliability. None of the existing work demonstrates high reliability of their genre annotation via inter-annotator agreement or presents a clear annotation procedure that is then proven to lead to a reliably annotated corpus.

The reasons for this differ. Corpora such as SANTINIS, KI-O4 and Syracuse have been annotated by a single person. As a result, their inter-annotator agreement measures cannot be computed. Given that SANTINIS and KI-04 explicitly searched for prototypical examples of a small set of categories, it is possible that the annotation could be recreated by several annotators but it cannot be assured and there are no publicly available guidelines to test. The MGC, I-EN-Sample and KRYS I corpora have been double-annotated. However, agreement measures were low ($\alpha = 0.56$ for MGC and $\alpha = 0.55$ for I-EN-Sample) as discussed in detail in Sharoff et al. (2010).[4] Table 3 shows the low percentage agreement for the KRYS I corpus in percentage agreement—chance-corrected agreement tends to be even lower.

The corpus constructed in Egbert et al. (2013, 2015) is annotated via crowd-sourcing. Four annotations were assigned to each web page via the crowd-sourcing website Amazon Mechanical Turk. However, reliability results are not high: on the eight main functional genres, on only 63 % of the web pages at least 3 out of four annotators are in agreement; for the fine-grained genres, on only 43 % of the web pages at least 3 out of 4 annotators are in agreement [see the pilot study in Egbert and Biber (2013)]. In Egbert et al. (2015) chance-corrected agreement is computed at a kappa of 0.47 and 0.40 for coarse- and fine-grained categories respectively, again showing only moderate agreement.

Overall, it is interesting that granularity is an insufficient explanation for low reliability results as in many corpora (coarse-grained categories in Egbert and Biber (2013), Egbert et al. (2015), I-EN Sample, MGC) reliability is low even for a relatively small (<20) number of categories.

Corpus design and expert annotation. There are two other issues regarding annotation in current corpora.

Firstly, many of these corpora are designed, i.e. constructed by a focused search for pages that are likely to fit a given category.[5] This is advantageous for the first test of an annotation scheme as one avoids noisy pages or borderline cases. Learning

[4] We refer the reader to Artstein and Poesio (2008) for a comprehensive survey of inter-coder agreement measures such as percentage agreement as well as chance-corrected agreements α and κ.

[5] The exceptions are MGC, which contains both a focused-search and a random web page collection, as well as I-EN and Egbert et al. (2013, 2015), which consist of random pages.

from prototypical examples can also be good for training automatic genre identification algorithms. However, it is unclear how manual or automatic results transfer to arbitrary web pages. In fact, Sharoff et al. (2010) show that human agreement tends to be even lower for arbitrary web pages than for web pages collected by focused search. A similar point is made by Rehm et al. (2008) who propose a designed corpus as a first step, with a corpus consisting of more randomly selected web pages as a second one. Unfortunately, the authors did not actually follow up with their own web genre corpus following this suggestion. In this paper, we remedy this gap.

Secondly, expert annotation can mislead with regards to the general applicability of the annotation scheme, especially if the same experts conducted annotation and developed the scheme (Riezler 2014). This was the case in SANTINIS, MGC and KI-04, for example. To avoid this problem, we use crowd-sourcing with a larger number of naive annotators that are distinct from scheme developers. In contrast to Egbert et al. (2013, 2015), who also use crowd-sourcing, we do not focus on coverage but on reliability, so that these efforts are complementary. To the best of our knowledge, this is therefore the first crowd-sourcing effort for genre annotation with demonstrably high inter-annotator agreement.

Size. Many existing collections are not large enough to ensure representativeness of genre classes. Table 2 shows the maximum, minimum and median number of web pages per genre category. As can be seen, they often have few annotated web pages *per category*, especially the KRYS-I and Syracuse corpora, while machine learning algorithms often require a reasonable number of training examples in order to produce satisfactory results. A notable exception is Egbert et al. (2015): although it also contains many genres with few or no examples, 24 of the 56 genres used are represented by over 100 pages.

Format. Another major drawback of some existing corpora is that they have been preserved in non-HTML formats such as PDF or plain text. For instance, each web page in KRYS I corpus is saved in PDF format. As a result, automated tools are needed to convert PDF to plain text or HTML format. However, these tools are error prone: therefore, some information may be lost or wrongly converted. In addition, previous studies in AGI show that HTML tags can improve the accuracy of genre classification (Kanaris and Stamatatos 2009) and should therefore be kept when collecting web genre corpora.

Topic diversity. There are genres which have a natural, strong correlation to certain topics, for example the genre label *recipe* has a clear connection to the topic label *food*. These types of correlations between genres and topics are true and explicit connections and will always exist. However, in some existing genre-annotated corpora, there are a number of correlations between genres and topics which are spurious in that they are due to the way the search for genre texts was conducted. For example, a large sample of the *frequently asked questions* texts in Santinis corpus (Santini 2007) come from web sites about *hurricanes*. Such spurious correlations can mislead investigations into typical genre properties—(Petrenz and Webber 2011), for example, show that the often best-performing bag-

Table 2 This table summarizes some characteristics of genre-annotated corpora

Corpus	Number of		Number of pages per genre			Format	Collection method	Reliability
	Pages	Genres	Min	Max	Median			
KRYS I (Berninger et al. 2008)	6200	10/70	6	117	97	PDF	Focused search	a.p.a.= 50.38 % (Table 3)
MGC (Vidulin et al. 2007)	1536	20	55	227	77	HTML with images	Both random selection and focused search	Low $\alpha = 0.56$ Sharoff et al. (2010)
HGC (Stubbe and Ringlstetter 2007)	1280	7/32	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 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 α=0.55 Sharoff et al. (2010)
Syracuse (Crowston et al. 2011)	3027	292	1	174	3	HTML only	Focused search	Not measured
The corpus in Egbert and Biber (2013)	1000	8/56	0	99	1	HTML	Random selection	43 % of the time three out of four annotations agreed
The corpus in Egbert et al. (2015)	53,000	8/56	0	3500	?	HTML	Random	Low κ 0.40

Question mark which stands for "unknown"

It illustrates low or not-measured agreement, the sometimes low number of web pages per individual genres, information loss by not preserving HTML in some corpora and the focus on designed corpora. If the genre labels are hierarchical, the statistics and agreement given are on the lowest level of the hierarchy. a.p.a. stands for average percentage agreement. See Artstein and Poesio (2008) for a comprehensive survey on inter-coder agreement measures such as percentage agreement, α and κ

Table 3 Human agreement for the KRYS I corpus (Berninger et al. 2008) which has *seventy* genre classes

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

Results illustrate a low percentage agreement

of-words features in AGI perform considerably worse when topic is varied. Therefore, AGI based on these features potentially learns topics rather than genres.

As far as we know, there is no corpus construction approach that explicitly looks into topic diversity of the resulting corpus. We propose a method how to approach this and discuss source and topic diversity explicitly.

4 Aims of this study: creating a reliable genre-annotated corpus via crowd-sourcing

Currently, there is no web genre annotation method established that results in demonstrably high inter-annotator agreement. We try to remedy this gap by building the Leeds Web Genre Corpus (LWGC) which fulfills the following criteria:

- It is reliably annotated for genre as measured by chance-corrected agreement. Reliability has currently been established for 15 genre classes. We also discuss extensibility of our procedure to other genre classes in Sect. 7.5.
- It avoids circularity by crowd-sourcing naive annotators that were not involved in annotation scheme development (Riezler 2014).
- Web pages have been saved in HTML format. Also, the appearance of each web page has been preserved by taking a screen shot of its whole content. The latter can facilitate using visual features as well as textual and HTML features in AGI.
- It contains a sub-corpus (LWGC-B) that used focused search to create a corpus with a substantial number of web pages for each individual genre category. LWGC-B has been collected from a diverse range of sources in order to avoid creating false correlations between genres and topics. We discuss an approach to measure topic diversity for genre corpora.
- It also contains a sub-corpus (LWGC-R) that approximates random web page collection to test (1) the transferability of the developed annotation scheme to arbitrary web pages and (2) to explore coverage of the current inventory of genre classes.

5 Genre inventory

The quality of manual annotations depends on the use of precise and consistent guidelines which include category definitions. Therefore, the development of the annotation guidelines must be seen as one of the crucial tasks in annotation projects. Although the main focus of this work is *not* the development of a comprehensive genre taxonomy, we still need clearly defined categories that our naive annotators have a chance of annotating with little training.

We used several criteria that all our genre classes needed to fulfill.

Form and function. First, we want to use only genre classes and terms that include form constraints in addition to functional constraints. This is in line with the definition we outlined in Sect. 2, and mirrors also (Kessler et al. 1997) who emphasize that a genre should not be so broad that the texts belonging to it do not share any distinguishing properties.

> we would probably not use the term genre to describe merely the class of texts that have the objective of persuading someone to do something, since that class (which would include editorials, sermons, prayers, advertisements, and so forth) has no distinguishing formal properties. (Kessler et al. 1997, p. 33)

Therefore, our genre inventory will automatically exclude the use of, for example, the broad register classes used in Egbert et al. (2013, 2015). We think it is quite possible that some of the broad, functional categories in previous annotation schemes led to low inter-coder agreement—an example are categories such as *informative* and *entertainment* in the MGC corpus (Vidulin et al. 2007) or functional genre categories as in I-EN Sharoff (2010) and Egbert and Biber (2013). Defining broad genre categories not only could cause disagreement between annotators, but it could also have a negative impact on automatic genre classification.

Common usage. For naive annotators we want to use genre names which they might have heard before and that are in common use (such as *forum*) and avoid expert linguistic terminology while remaining specific. This is not just a choice of convenience but also mirrors the fact that genres should be socially recognisable as postulated by the definition we give in Sect. 2.

Text orientation. As another constraint, we were interested in textual genres only and excluded all genres that are mainly visual or include little text (such as link lists, web pages with just a video or a series of pictures etc.).

Variety of different functions. Although our genre names and descriptions will include both form and function, we want to include genres that cover a broad range of functions or what (Biber 1991) calls text type dimensions. Thus, we want genres from the narrative as well as the non-narrative spectrum or from the colloquial/spontaneous vs. edited text spectrum.

Limited set. As this was our first study on genre annotation via naive users, we decided to start with a limited genre palette instead of a complete taxonomy. We therefore made a list of all previously used genre terms that fulfilled the criteria above, mapping equivalent terms as best we could, and chose a subset of 15 genres

Table 4 Definition of genre labels in the LWGC

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): Instruction/How to (instruction)	Web pages created with intention to sell 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): Editorial	A report of recent events 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): Frequently Asked Questions (faq)	A detailed description of someone's life questions commonly asked about a particular topic, in list form
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

To save space, in this paper we use the abbreviations of genre labels which are specified after the genre names

from a wide spectrum of form and function. We also focused on genres that we hypothesized to be frequent on the Web due to our own informal experience such as blogs, news articles or forums.

In addition, we also tried to narrow our definitions down as much as possible while staying with socially recognizable forms: this led for example, to the inclusion of the genre *recipe* as distinct from other how-to instructions. We think that this actually allows the definition of other how-tos to be more precise. In Sect. 6.5 we will show that, in accordance with our intuition, the genre recipe is indeed distinct from other instructions with regard to length and type/token distributions.

Final set. Table 4 shows the set of 15 genre labels and their definitions. We are fully aware that our set of criteria could also lead to a set of different genres: however, this set of genres will allow us to test crowd-sourcing for a wide variety of forms and functions and includes many web-typical genres, such as homepages and forums. Other approaches can use their own genre palettes as long as they fulfil the

Table 5 This Table illustrates which genre classes in our corpus are also included in existing genre-annotated corpora

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

same criteria and have reasonable hope that a similarly designed crowd-sourcing effort will also lead to good annotation for them.

Table 5 shows how these 15 selected genre classes correspond to those used in other genre-annotated corpora. However, since different genre-annotated corpora used different genre classes with different levels of granularity, any one-to-one comparison between our genre labels and their genre classes can only be approximate. For example, the genre label *journalistic* in MGC can include several genres in our corpus such as *news*, *editorial*, *interviews* and *reviews*. Another example is the *periodicals* (newspaper, magazine) category from the KRYS I corpus which is very broad and can include many genre classes such as *recipe*, *interview* and *reviews*.

The genre inventory in Table 4 applies to both sub-corpora of the LWGC. We explore the coverage of our scheme in Sect. 7.

6 LWGC-B: a web genre corpus designed via focused search

Web corpora are categorized into designed and random corpora according to their collection method (Kilgarriff 2012). The content of a designed corpus is selected based on its design specification, normally following a focused search method. In contrast, the content of a random corpus represents a (more or less faithful) snapshot of the web. HGC (Stubbe and Ringlstetter 2007) and UKWac (Baroni et al. 2009) are examples of designed and random corpora, respectively.

As explained in Sects. 1 and 3, we use a designed corpus as the first step for testing our annotation scheme and crowd-sourcing effort, for two reasons. First, we can provide a corpus with a large number of web pages for each category via this method. While collecting random web pages is fast and cheap, there is no guarantee that it fulfills this criterion. Second, manually collected, prototypical examples provide a good test bed for using naive annotators. If agreement cannot be established on the prototypical pages, it is unlikely to be achieved on random pages. It is also possible that prototypical examples are better for training machine learners. The use of a designed corpus was also suggested by Rehm et al. (2008) as an initial step when building a reference corpus of web genres.

On the flip side, a designed corpus will not give us an accurate representation of the actual genre distribution on the web nor will it tell us the coverage of our annotation scheme. Annotation results on clear and prototypical web pages are also likely to overestimate inter-annotator agreement (Sharoff 2010). We will investigate those issues in Sect. 7 where we collate and annotate a smaller, random corpus, the LWGC-R.

6.1 LWGC-B: corpus compilation

We hand-selected web pages mainly from existing web directories, particularly the Yahoo Directory[6] and Open Directory Project[7] websites. We selected 3964 web pages from a diverse range of sources to avoid creating false correlations between topic and genre labels. We will discuss the source and topic diversity of the corpus further in Sect. 6.6.

In the next phase, we used the KrdWrd tool (Steger and Stemle 2009) to download the web pages in HTML format. However, only saving a web page in HTML format does not guarantee the preservation of its appearance. To achieve this aim, we could, for each web page, save all its graphics and style files, or take a screen shot of its whole content. We chose the second option and used KrdWrd to also preserve each web page as an image.

6.2 LWGC-B: annotation procedure

After collection, the corpus needs to be annotated with the set of chosen genre labels (see Sect. 5), which can be a very time consuming and expensive task. However, in recent years, the advent of crowd-sourcing (e.g. via Amazon Mechanical Turk[8]) has facilitated annotation tasks so that this phase can be done more cheaply and faster than ever before. Amazon Mechanical Turk (MTurk) has been used for a variety of labelling and annotation tasks in Natural Language Processing e.g. word sense disambiguation, word similarity, text alignment, temporal ordering (Snow et al. 2008); machine translation (Callison-Burch 2009) and building a question answering dataset (Kaisser et al. 2008). It has also been used for genre annotation by

[6] http://dir.yahoo.com/.

[7] http://www.dmoz.org/.

[8] https://www.mturk.com/mturk/welcome.

Egbert et al. (2013, 2015) but without establishing high inter-annotator agreement (see Sect. 3).

In addition to saving expense and time, we can ensure easy re-use of the annotation scheme if even naive annotators with short guidelines achieve high reliability. The fact that the annotators are independent of scheme developers also avoids circularity in annotation (Riezler 2014).

6.2.1 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 carried out by untrained MTurk workers (turkers) all around the world for a small amount of money. The main advantages of Mturk are low cost and speedy task completion as well as its infrastructure, which allows the requesters to develop their HITs using standard HTML and Javascript.

With turkers, quality control is crucial in order to detect poor quality or randomly selected answers. Moreover, Mturk HITs, like any other web-based interface, are vulnerable to automated scripts, also known as bots, which are used by some turkers in order to maximize their income (Mason and Suri 2012). We therefore used two types of qualification criteria in our HIT design, as provided by MTurk.

Firstly, MTurk provides "system qualifications," which are independent of the specific task created. They include HIT submission rate (the percentage of accepted HITs eventually submitted by the turker), HIT approval rate (ratio of HITs approved by the requester compared to the total number of HITs submitted by the turker), HIT rejection rate (ratio of rejected HITs compared to the total number of HITs submitted by the turker) and location (the worker's country of residence).

The second type of quality control measures is task-specific. It includes the possibility of a pre-task qualification test designed by the requesters. Up to five qualification criteria can be assigned to a HIT by the requester. Only turkers who pass these qualification measures are permitted to complete the HITs. With regards to after-task quality control, Mturk enables the requesters to download and (automatically or manually) review the submitted works, then reject poor quality data and only pay for the HITs which they approve. In the next section, we describe both the system qualifications and task-specific pre- and after-task quality controls that we use.

6.2.2 HIT design and quality control

This section describes the details of of our HIT design and quality control measures.

HIT design. Turkers were presented with the list of our 15 genre categories together with short guidelines that allowed them to view category definitions (see Table 4). They were also able to view example pages for the categories, if wished for. As our genre inventory is not exhaustive, annotators were also allowed to choose the option *other* for web pages that do not fit any of the 15 classes. In order

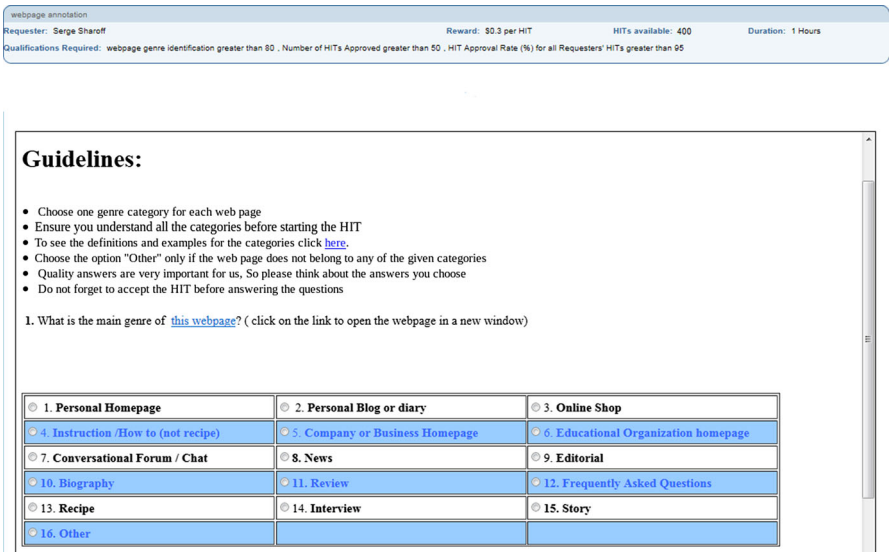

Fig. 1 Screen-shot of genre annotation task on Mturk website

to keep the annotation task simple, we decided to choose the single-labeling method, i.e. each web page could only receive a single genre label, despite the fact that there are some web pages that might belong to more than one genre class (Crowston and Kwasnik 2004; Kessler et al. 1997; Santini 2008). Annotators needed to click on a link to open the web page to be annotated—the cached web page would then open in a separate window. Figure 1 shows a screen-shot of the annotation task.

A single HIT includes 10 web pages to be annotated, both as this is more time and cost-effective, and because we are going to use this feature in quality control as described below.

Quality control. With regard to system qualifications, we restricted the range of workers who can complete our task. As we were looking for experienced workers, we only allowed workers who had successfully completed at least fifty HITS previously. To ensure diligence, we restricted the task to workers with an approval rate of 95 % or greater.

As a task-specific pre-task qualification test, we let turkers read the definitions and examples of genre classes and then complete a trial HIT of ten genre annotations on pages that we deemed highly prototypical and therefore should be annotated without much scope for error. Only turkers who completed this qualification test with a score of at least 80 % were allowed to take part. This was supposed to weed out bots and random clickers.

For after-task quality control without excessive manual work or introducing substantial expert bias, we used one of the ten web pages to be annotated per HIT as a "trap" question. We selected a set of twenty web pages that the first author of this paper judged as unambiguous and clear examples of one of our predefined genre

Table 6 Landis and Koch interpretations (Landis and Koch 1977) of Fleiss's kappa (Fleiss 1971)

Fleiss's kappa (Fleiss 1971)	Level of agreement
<0	Poor
0.0–0.2	Slight
0.2–0.4	Fair
0.4–0.6	Moderate
0.6–0.8	Substantial
0.8–1.0	Perfect

categories. We used these web pages as trap questions. We performed semi-automated monitoring of the annotations by checking the answers to the trap questions and rejected the workers who did not give the right answers to the trap questions at least 80 % of the time.

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: Snow et al. (2008) compared the quality of annotation done by experts and Mturk workers and concluded that an average of 4 turkers often provides expert-level label quality.

6.3 Inter-coder agreement measures

In Natural Language Processing and machine learning, a reliably annotated dataset plays a crucial role. The results of research based on unreliable annotation can be considered as untrustworthy, doubtful and even meaningless. In order to measure the reliability of annotation, different annotators judge the same data and the inter-coder agreement is calculated for their judgments. The most commonly used inter-coder agreement measures are: percentage agreement, S (Bennett et al. 1954), Scott's π (Scott 1955), Cohen's or Fleiss κ (Cohen 1960; Fleiss 1971) and Krippendorff's α (Krippendorff 1970) [see Artstein and Poesio (2008) for a comprehensive survey of inter-coder agreement measures].

Percentage or observed agreement is the simplest measure of agreement among coders. However, this measure does not take into account agreement which is expected to happen by chance. As a result, it can overestimate true agreement. Therefore, other inter-coder agreement measures which correct for chance agreement must be computed. Originally these coefficients (such as Scott's pi π and Cohen's kappa κ) were proposed for calculating inter-coder agreement between two annotators. Then Fleiss (1971) proposed a generalization for Scott's π (called Fleiss' κ) and Davies and Fleiss (1982) one for Cohen's κ. Although these two measures often have very similar values, there is one crucial difference between them. For calculating expected agreement for Scott's π and Fleiss' κ, we only take into account the combined judgments of all coders and not the number of items assigned to each category by each individual coder. In contrast, for calculating

Table 7 Inter-coder agreement for individual categories in LWGC-B shows substantial agreement among the coders. Therefore, annotations for all the genre classes are highly reliable

Genre labels	Percentage	Fleiss's κ agreement
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

expected agreement for Cohen's κ, we take into account the number of times each individual coder assigns an item to a category.

Since in Mturk the annotations have been done by various workers, Cohen's κ is not applicable as it needs a consistent set of annotators for all items. Therefore, like other annotation studies using crowd-sourcing (Mohammad and Turney 2012; McCreadie et al. 2011; Bentivogli et al. 2011), we calculated Fleiss's kappa (Fleiss 1971) for the annotation. The next section presents inter-coder agreement results.

6.4 LWGC-B: annotation study results

Overall, 42 turkers participated in annotating the corpus. The annotation task was completed within seven days for a total cost of $820. We paid 40 cents per HIT, therefore 4 cents per page to be annotated (a HIT included 10 pages).

We achieved high reliability with a percentage agreement of 88.2 % and Fleiss's kappa of 0.874. Based on the interpretation of the inter-coder agreement value by Landis and Koch (1977) (Table 6), this value shows perfect agreement between the annotators.

We also computed Fleiss's kappa for each single category in order to identify the most and the least agreed-on genre classes. To compute single category κ for a target category t, we merge all other categories into one $non - t$ category and then compute agreement between t and $non - t$. Table 7 shows the inter-coder agreement for individual genre classes. κ values for the individual categories illustrate substantial agreement among the coders for all categories and, as a result,

Table 8 Distribution of different types of inter-annotator agreement in the LWGC-B

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

annotations for all the genre classes are highly reliable. The category *recipe* was the easiest one for the annotators to identify whereas *company/ business home pages* caused the most disagreement (this genre category was mostly confused with *shop*).

The next phase of building a reliable genre annotated dataset is to convert the annotated corpus into a gold standard. There are a number of different methods to do so (Beigman Klebanov and Beigman 2009). For instance, the annotators can discuss together to reach agreement on the disagreed items (Litman et al. 2006) or if more than two annotators engage in the annotation task, a majority vote approach can be employed (Vieira and Poesio 2000). 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 the instances which cause disagreement can be excluded from the dataset (Beigman Klebanov and Beigman 2009).

As we employed Mturk for annotation, reaching agreement through discussion between annotators is not possible. We also decided against expert labelling as we still wanted to keep involvement of the annotation scheme developers to a minimum. 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.[9]

In order to analyze how often the annotators agreed with each other, we calculated the percentage of each type of inter-annotator agreement (Table 8). For more than 74 % of the web pages all five annotators agreed and for 95 % of the data at least four annotators agreed on a single label, indicating high level of agreement between the coders. Low percentage of the other five types of inter-coder agreement confirms the high value of κ 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 a majority vote for eight web pages, the final labels for these instances were assigned by the first author of this paper.

[9] For example, 3, 1, 1 indicates 3 annotators agreeing on a category x whereas the fourth and fifth annotator choose category y and z, respectively.

Table 9 The corpus statistics for LWGC-B

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 per category	266
Number of tokens	7,205,820
Number of types	130,254
Number of sentences	329,861

Table 10 Text statistics for individual categories in the LWGC-B

Genre	Number of											
	Sentences			Tokens			Types			Types/token ratio		
	Max	Min	Med	Max	Min	Med	Max	Min	Med	Max	Min	Med
php	326	0	11	4,165	21	241	1,232	17	142	1.00	0.22	0.57
com	195	0	11	4,906	32	330	1,390	29	172	0.90	0.22	0.53
edu	179	0	10	5,501	12	396	1,960	11	209	0.93	0.13	0.54
blog	1,041	14	139	19,488	214	2,905	2,926	141	882	0.69	0.09	0.30
shop	600	0	33	25,651	71	1337	7,459	45	456	0.69	0.05	0.33
instruction	595	11	99	12,767	199	1,219	1,988	102	447	0.57	0.14	0.36
recipe	584	2	20	11,445	123	428	2,218	68	221	0.74	0.15	0.50
news	702	7	41	16,642	271	1312	3,052	140	603	0.64	0.15	0.45
editorial	511	9	45	10,537	311	1367	2,309	187	661	0.63	0.19	0.47
forum	619	2	60	13,010	269	1454	1,932	144	473	0.60	0.12	0.32
bio	2,465	4	67	23,838	198	1616	4,603	103	625	0.67	0.15	0.39
faq	613	5	54	14,312	119	971	2,220	68	355	0.73	0.13	0.36
review	1,107	12	96	19,261	174	2,094	2,979	118	634	0.73	0.15	0.31
story	1,012	10	98	10,521	239	1777	2,043	98	586	0.56	0.16	0.33
interview	1,243	29	150	19,687	380	2,487	3,601	153	733	0.50	0.13	0.30

Max, min and med are abbreviations of minimum, maximum and median, respectively

6.5 LWGC-B: 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 (see Table 9). The corpus consists of 3964 web pages, distributed across 15 genres.[10] Each genre is represented by at least 184 web pages. The distribution is pretty balanced between the genres as we intended for this part of the corpus. It

[10] Although individual annotators used the label *other*, it was never the majority annotation due to the focused search collection.

Table 11 Statistics for individual categories in the LWGC-B illustrating source diversity of the corpus

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

Max, min and med are abbreviations of minimum, maximum and median, respectively

Table 12 Contingency table for calculating log-likelihood

	The web page	Whole corpus
Freq of word	a	b
Freq of other words	c–a	d–b

a and *b* are the frequency of the word in the web page and the whole corpus, respectively. *c* corresponds to the number of the words in the web page and *d* is the number of the words in the whole corpus

contains more than 7 million words which makes it approximately seven times bigger than the Brown corpus.

Table 10 compares genre classes in terms of text statistics. A number of interesting observations can be made from the individual categories' statistics. First, the length of an average home page is less than for most other genre categories in this corpus. On the other hand, *personal blog* and *interviews* contain the longest texts. A closer look at the corpus statistics also reveals that home pages tend to have high type/token ratio compared to other categories. Recipes are substantially shorter than other types of instructions. Based on these observations, it seems that automatic genre classification algorithms could benefit from the discrimination power of these statistics as features.

In order to investigate how up-to-date our corpus is we approximated on which dates the web pages were published or last modified. We used the stanford named entity recognizer (Finkel et al. 2005) to identify all the dates in each page. The most

recent date was taken as the publish or last modified date. The results show that about 75 % of the pages were last updated in the years 2010–2012.

6.6 LWGC-B: investigating source and topic diversity

Collecting data for a genre category from topically similar sources was one of the drawbacks of some of the existing genre-annotated corpora mentioned in Sect. 3. In the construction of our corpus, we therefore tried to compile web pages from a diverse range of sources. We calculated source-diversity statistics for each genre in Table 11.

We can see that our focused search avoided collecting too many web pages per site as most genre categories have a median of one web page collected per site. This is positive as it avoids associating genres with specific websites and layouts which are subject to fast change (although, of course, genres also change over time). However, there are still some web sites that might be over-represented such as the maximum of 23 pages from a single shopping web site (which was Amazon).

As even different sources could be on the same topic, we conducted an additional investigation into the topic diversity of our corpus by extracting and comparing keywords of web pages in each genre category. The underlying assumption of this approach is that if web pages in a genre category have topically similar keywords, then that category is not represented by a sufficient variety of topics. We used the log-likelihood statistic (Dunning 1993) to identify words of a web page which have a significantly higher frequency in that page than in the whole corpus. The keyword extraction procedure consists of the following steps (Rayson and Garside 2000):

1. We produced a word frequency list for each web page as well as the whole corpus.
2. For each word in the word frequency list for each web page, we calculate the log-likelihood statistic by constructing the contingency table shown in Table 12, where a and b are the frequency of the word in the web page and the whole corpus respectively; c corresponds to the number of the words in the web page and d is the number of the words in the whole corpus. We can compute the log-likelihood value based on this formula:

$$LL = 2\left(\left(a\log\left(\frac{a}{E1}\right)\right) + \left(b\log\left(\frac{b}{E2}\right)\right)\right) \quad (1)$$

   where $E1 = c\frac{(a+b)}{(c+d)}$ and $E2 = d\frac{(a+b)}{(c+d)}$.
3. Then we sort the word frequency list of each web page according to their LL values. The words with the highest LL values are the keywords of the web page as they occur more frequently in the page than in the whole corpus (when normalized for page/corpus size).

We only considered keywords which are significant at the level of $p < 0.0001$ and also removed some common words such as pronouns and determiners. Next, we needed to generalize from individual web pages to genre classes. To do so, we counted the number of web pages in each genre class that a keyword appears in.

Table 13 Keywords from genre categories in LWGC-B

faq (201)		blog (244)		com (264)		editorial (310)		edu (299)	
58	can	70	posted	54	company	93	opinion	130	school
51	questions	50	january	48	services	69	news	124	students
46	information	47	comments	33	products	59	editorial	99	university
45	do	46	blog	25	service	50	blogs	71	campus
44	are	43	was	23	ltd	44	state	69	research
33	does	34	december	20	systems	41	columns	66	student
32	how	31	labels	20	corporation	39	autos	43	programs
29	frequently	31	day	19	clients	38	editorials	43	college
28	services	28	but	18	solutions	33	obituaries	43	academic
26	is	27	august	16	website	31	local	40	undergraduate
26	if	25	share	16	management	31	business	40	events
25	may	25	had	16	contact	29	columnists	37	faculty
24	what	24	july	16	construction	29	city	35	international
23	was	24	christmas	15	design	29	cars	35	graduate
22	will	23	twitter	15	business	29	ads	35	alumni
22	program	23	just	13	provide	28	jobs	33	admissions
22	available	23	april	13	group	27	reprints	32	high
21	top	23	am	13	corporate	27	government	31	learning
21	site	23	about	12	support	26	*obama*	31	information
20	page	22	october	12	industry	26	editor	31	education
bio(242)		forum (280)		instruction (231)		interview (185)		news (330)	
62	biography	201	posts	102	how	81	do	135	news
39	became	164	forum	46	step	77	was	104	said
27	had	143	join	43	or	74	did	30	police
26	*music*	137	thread	43	if	57	what	30	latest
25	later	135	date	41	do	41	think	29	photos
24	will	105	location	37	will	38	were	28	headlines
24	father	102	member	35	make	35	they	26	tuesday
24	as	93	pm	29	use	35	me	26	government
23	have	92	reply	28	was	34	people	25	sport
23	died	82	quote	26	can	34	had	25	minister
21	published	68	am	25	tips	32	because	25	health
21	born	67	post	24	are	30	really	25	blogs
20	married	66	profile	23	get	29	interview	24	sports
20	life	60	view	21	yourself	28	music	23	watch
20	film	60	forums	20	paper	28	lot	23	sun
20	during	57	re	20	job	27	like	23	national
20	award	53	thanks	19	water	26	there	23	former

Table 13 continued

bio(242)		forum (280)		instruction (231)		interview (185)		news (330)	
20	*album*	46	replies	19	need	26	know	22	search
19	children	43	hi	19	instructions	24	would	22	*president*
18	were	38	linkback	19	be	23	just	22	lifestyle
php (304)		**recipe (332)**		**review (266)**		**shop (292)**		**story (184)**	
38	*research*	145	recipes	126	review	89	price	72	said
30	university	139	recipe	91	reviews	79	accessories	37	then
23	website	90	cup	39	product	65	shop	34	could
18	cv	75	sauce	38	very	65	shipping	33	old
16	site	69	cooking	37	rating	64	product	33	little
16	guestbook	67	garlic	35	recommend	61	free	31	shall
14	welcome	64	butter	34	service	57	more	31	came
14	page	63	pepper	33	comment	54	items	30	eyes
12	*economics*	62	sugar	31	overall	49	*amazon*	29	door
12	blog	60	the	27	helpful	47	reviews	28	words
11	teaching	60	ingredients	27	*book*	47	*clothing*	28	*king*
11	publications	60	cheese	26	great	46	delivery	27	thought
11	*professor*	58	add	25	excellent	46	buy	25	stood
11	pdf	56	teaspoon	25	but	45	customer	25	went
10	social	56	cook	24	*video*	43	gift	25	replied
10	*engineering*	55	onion	24	useful	42	products	25	man
9	web	55	*chicken*	24	reviewer	41	see	25	looked
9	projects	54	minutes	24	good	41	basket	24	cried
9	personal	53	chopped	23	out	40	store	24	woman

Italics indicate keywords that are likely topic-specific

Table 13 shows the keywords which appear in the highest number of web pages for each genre category of our corpus. Each number shows the number of documents that the corresponding word has been selected as a keyword for. Although to a certain degree subjective, we indicated potentially spurious "topic invasion" in our corpus with italics in the Table.

A qualitative analysis of the results presented in Table 13 shows very few topic-specific words. As wished for, the majority of the words are genre-specific. For example, frequently asked questions are not distinguished by keywords that indicate FAQs on a specific topic but instead by general question words (such as *how* or *what*) and parts of the genre name itself. An exception is the keyword *program* which might indicate several FAQs on programming languages. Similarly, blogs and forums are not distinguished by specific topics but by, for example, posting dates for blogs, and forum-specific words such as *member, join, thread*. An

Table 14 Keywords from some of the genre categories of the existing web genre corpora

faq (200)		php (200)		blog (200)		shop (200)	
SANTINIS (Santini 2007)							
110	*hurricane*	26	*math*	40	but	32	click
109	*noaa*	16	page	39	march	29	*dvd*
107	center	16	*mathematics*	30	just	28	price
84	*aoml*	15	university	29	posted	26	more
65	*tropical*	13	unl	28	had	25	basket
57	*tax*	12	lincoln	28	comments	22	uk
48	publication	12	guestbook	28	blog	22	info
47	faq	12	dk	27	like	21	delivery
46	form	11	*research*	25	am	21	add
42	pdf	11	*mathematical*	24	get	17	order
41	references	10	teaching	22	february	17	here
40	*cyclones*	10	bradley	20	know	17	details
37	*income*	9	theory	20	got	16	save
36	topic	9	office	20	going	15	summer
33	file	9	*nebraska*	19	really	15	offers
32	back	9	homepage	18	trackback	14	*games*
31	return	8	thesis	18	think	14	free
31	if	8	edu	18	there	14	*flowers*
26	*storm*	8	department	18	as	14	catalogue
25	*weather*	7	*mit*	17	very	13	product
php (126)		help (139)		shop (167)		forum (127)	
KI-04 (Meyer zu Eissen and Stein 2004)							
19	*intelligence*	52	do	42	store	41	post
18	computer	47	how	30	price	41	pm
17	*research*	40	what	23	cart	39	forum
15	*artificial*	37	can	22	shop	39	am
11	proceedings	36	faq	19	*books*	37	posts
11	conference	33	if	17	shipping	37	message
10	systems	26	there	17	gifts	30	reply
10	science	24	search	16	gift	24	thread
10	*reasoning*	24	com	16	buy	23	topic
10	homepage	23	be	15	products	23	forums
9	*professor*	23	web	15	click	22	posted
8	computational	22	site	14	more	20	view
7	member	22	help	14	*book*	20	quote
7	*engineering*	22	file	13	*music*	20	new
7	dr	21	will	12	sellers	19	re
7	*ai*	20	why	12	here	18	to
7	*aaai*	20	this	11	valentine	18	send

Table 14 continued

php (126)		help (139)		shop (167)		forum (127)	
6	*simulation*	20	server	11	top	18	profile
6	publications	18	use	11	sale	17	last
6	language	18	http	11	now	17	edit
blog (77)		forum (82)		faq (70)		fiction (67)	
MGC (Multi-labelled Genre Collection) (Vidulin et al. 2007)							
30	posted	29	posts	34	can	36	had
22	pm	20	reply	31	if	35	said
20	blog	20	message	28	was	24	back
18	comments	19	quote	28	what	23	up
15	am	18	thread	24	how	22	looked
13	blogs	18	pm	24	do	19	eyes
11	but	17	am	24	faq	18	down
10	weblog	16	profile	23	are	18	could
10	people	16	post	19	http	18	would
10	comment	15	send	17	version	17	then
9	trackback	12	private	17	use	17	out
9	like	12	posted	17	q	17	into
9	here	11	view	16	html	16	door
9	april	11	topic	16	file	16	but
8	your	11	offline	15	be	15	which
8	think	11	list	15	using	15	room
8	site	11	forum	15	user	15	man
8	october	11	buddy	14	does	15	just
8	march	10	mode	14	com	15	head
7	will	10	may	14	web	14	felt

Italics indicate keywords that are likely topic-specific

exception is the genre category *recipe* where an unavoidable correlation to the topic *food* holds. Even there our corpus did not contain only recipes of a specific type, such as mostly vegetable recipes—instead keywords indicate flexible widely used ingredients (with the possible exception of *chicken*). Some potentially topic-dependent keywords such as *cars, autos* for editorials are not due to the corpus containing many editorials about cars but because of frequent advertising links in the boiler plate. It is also important to note that some topic-like keywords probably mirror the current distribution of web genres, such as the fact that many personal home pages are of scientists.[11]

[11] Note that the topics of the scientists' homepages are widespread with the most being about 3 % coming from engineering and economics each.

In order to compare the topic diversity of our corpus to prior work, we also extracted keywords from comparable genre classes in existing web genre corpora. Table 14 depicts some of the results. Qualitative analysis shows that the *faq* category in SANTINIS (Santini 2007) is the least topically diverse category. Almost all the web pages in this genre class are about *hurricane* and *tax*. Also, Table 14 shows that although keywords from categories such as *blog* and *forum* are mainly genre-specific, *personal home pages* in KI-04 (Meyer zu Eissen and Stein 2004) and SANTINIS seem to have too big a proportion from Artificial Intelligence researchers and mathematicians, respectively (over 10 % each).

7 LWGC-R: human annotation study on random web pages

So far we described different phases of constructing a designed web genre corpus. We chose to build a designed corpus as opposed to a random corpus because we wanted to have a balanced collection with a large number of web pages per genre category. The result of human annotation showed high inter-annotator agreement. The questions that we are seeking to answer in this section are twofold: firstly, can we achieve such high inter-annotator agreement on more arbitrary web pages, as well? Secondly, how good is the coverage of our genre inventory when applied to web pages that are not selected by focused search for particular genres?

In order to answer these questions, we repeated the same annotation study on a random corpus that builds on web search results. The following subsections describe the corpus collection, the corpus annotation and the results of the experiment in detail.

7.1 LWGC-R: web page collection

We use random conjunctive queries to a search engine for collecting an approximation of random web pages (see Manning et al. (2008, p398f.) for an in-depth discussion of the difficulties of collecting a random part of the Web). The BootCat toolkit (Baroni and Bernardini 2004) offers an easy way to use such random conjunctive queries via seed keywords.

Two things distinguish this method from a truly random web page collection (which would only be possible if we had access to a snapshot of the whole web). Firstly, if the queries are topic-specific such as *Rafael Nadal, tennis*, then we naturally will get topic-specific pages back. Therefore, we need to choose very general seeds in our case. We follow (Sharoff 2006) and use a list of the 500 most frequent words extracted from the BNC corpus as seeds. These are mostly function words. BootCat creates a list of *n*-tuples out of the seed words by randomly combining them. We used 3-tuples in this experiment (e.g., *have, we, which*). These 3-tuples are used as random conjunctive queries to a search engine. Secondly, as search engines, such as Google, rank and retrieve web pages based not only on keyword occurrence but also on their popularity, we actually do not get a truly random result either but rather a snapshot of popular web pages. In our case, this is not necessarily a disadvantage as being able to label the most used parts of the web

is important. However, there is also a genre bias when using the very top-most results which tend to be commercial home pages (Lim et al. 2005). Therefore, we ignored the first 30 URLs retrieved for each query and collected the 20 URLs which were ranked from 31th to 50th positions. Overall, fifty queries were sent to a search engine via BootCat, leading to the collection of 1000 URLs. After the URL collection phase, we downloaded the web pages using the KrdWrd tool (Steger and Stemle 2009).

We call this part of the corpus LWGC-R(andom). It must be noted that, even with our safeguards, the use of a search engine will still bias our corpus towards certain pages, in particular pages indexed by the search engine, more popular documents, as well as longer and recent documents (Manning et al. 2008, p398f.).

7.2 LWGC-R: annotation procedure

We carried out exactly the same annotation study as for LWGC-B, using Amazon Mechanical Turk. Annotators had the option to choose one of our 15 predefined genre categories or the option *other* for each web page. We set the number of annotations per web page to five. Moreover, the same quality control measures used in the experiment described in Sect. 6.2.2 (e.g. trap question, qualification test and high approval rate) were also adopted in this experiment. The annotation cost 222 Dollars.

7.3 LWGC-R: annotation results

To measure the reliability of the annotation, we calculated the inter-coder agreement measures. For this experiment, the percentage agreement is 78.15 % and κ is 0.712, which shows substantial agreement between the annotators (see Table 6). Therefore we can consider the annotation reliable.

We also calculated κ for individual genre labels (Table 15). The κ value is above 0.6 for all genre labels except *story* and *interview*. Quite importantly, the agreement for the category *other* is high which means that the current genres cannot only be easily delimited from each other (as in LWGC-B) but also from other, arbitrary, web pages. However, the κ value for the two genre classes *story* and *interview* is around zero, despite the fact that they have a very high observed or percentage agreement (99.9 and 99.8 %, respectively). A κ of around zero usually indicates very poor agreement. However, this interpretation of the chance-corrected agreement coefficient like π and κ only makes sense if the categories occur reasonably often (Feinstein and Cicchetti 1990). In contrast, the two categories *story* and *interview* were hardly ever chosen as can be seen in the fourth column of Table 15 where we indicate the number of times each category was chosen by the annotators. Due to the low number of samples of the two categories in the random corpus, we cannot draw definite conclusions with regard to the reliability of these two categories.

The comparison between the results of the annotation on the designed corpus LWGC-B and the random web pages in the LWGC-R reveals that the κ values on the more randomly selected web pages are lower. This could be due to two reasons: First, it could be because the random dataset is highly skewed. Second, it is harder to obtain a high inter-coder agreement for random web pages as these will include

Table 15 Inter-coder agreement for individual categories in LWGC-R shows substantial agreement among the coders

Genre labels	Percentage agreement	Fleiss's κ	N.T.C.A
Personal homepage	0.997	0.741	39
Company/Business homepage	0.888	0.646	961
Educational Organization homepage	0.993	0.707	64
Personal blog/Diary	0.979	0.611	83
Online shops	0.966	0.774	414
Instruction/How to	0.946	0.645	423
Recipe	0.999	0.928	43
News article	0.952	0.791	626
Editorial	0.991	0.667	67
Conversational forum	0.994	0.738	51
Biography	0.998	0.892	28
Frequently asked questions	0.993	0.757	58
Review	0.996	0.775	48
Story	0.999	−0.0004	2
Interview	0.998	−0.0008	4
Other	0.847	0.685	2089

N.T.C.A stands for number of times chosen by the annotators. For example, the category story has been chosen only two times by the annotators

more borderline or even hybrid cases. To provide more insight into this annotation study, we also compute the percentage of each type of inter-annotator agreement in Table 16. For 59.40 % of the web pages in LWGC-R all five annotators agreed and for more than 80 % of the data at least four annotators agreed which indicates high level of agreement between the coders. However, when we compare the two Tables 16 and 8 we see that annotators find it harder to agree on the random web pages. Nevertheless, the result of this study still shows substantial agreement between the annotators and, as a result, it was a successful annotation study.

We again employed the majority vote strategy to assign the final label to the disagreed web pages in this experiment just as for the designed corpus. As shown in Table 16, there are seven possible types of inter-annotator agreement when there are five annotators. However, there is no majority for the last three types. Therefore, as we did not have a majority vote for 34 web pages, we excluded them from the gold standard corpus.[12]

The distribution of the genre categories in LWGC-R is very skewed (Table 17). While genres such as *company home pages* and *news articles* comprise a high percentage of the total number of web pages in LWGC-R, other genre categories such as *biography* and *personal home pages* have very few web pages assigned to them. No web page represents the genres *story* and *interview*.

[12] This differs from the procedure in the designed corpus where for the eight pages without majority vote we used an expert label, instead. However, for several of the 34 web pages without a majority vote in the random corpus, the expert used (paper first author) was herself unsure of the label the page might belong to. Therefore, we excluded these pages from the gold standard.

Table 16 Distribution of different types of inter-annotator agreement in the LWGC-R

Types of inter-coder agreement	# of web pages	% of web pages
5,0	594	59.40
4,1	219	21.90
3,1,1	31	3.10
3,2	122	12.20
2,1,1,1	4	0.40
2,2,1	29	2.90
1,1,1,1,1	1	0.10

Table 17 Genre distribution in the LWGC-R

Category	# Web pages	% of the corpus
other	438	45.34
com	167	17.29
news	117	12.11
shop	79	8.18
how-to	76	7.87
blog	16	1.66
edu	12	1.24
editorial	12	1.24
faq	10	1.04
review	9	0.93
recipe	9	0.93
php	8	0.83
forum	8	0.83
bio	5	0.52
story	0	0
interview	0	0

7.4 LWGC-R: source and topic diversity

As noted in Sect. 3, a corpus used for automatic genre classification must be source and topic diverse. To achieve this for LWGC-B, we collected data from a wide range of sites. In contrast, the LWGC-R corpus was collected randomly, and it is interesting to see how topic and source diverse this corpus is.

We investigate the source diversity of the LWGC-R corpus by calculating the maximum, minimum and median number of websites per genre category (Table 18). The result shows that web page selection via random conjunctive queries as we used for LWGC-R collected data from a diverse range of websites. The maximum number of web pages selected from the same site is very low for all categories with the exception of the category *other*, where 31 web pages are selected

Table 18 Statistics for individual categories in the LWGC-R illustrate source diversity of the corpus

Genre	Number of		Number of pages from the same website		
	Web pages	Websites	Max	Min	Med
php	8	8	1	1	1
com	167	167	1	1	1
edu	12	12	1	1	1
blog	16	16	1	1	1
shop	79	66	7	1	1
instruction	76	69	5	1	1
recipe	9	9	1	1	1
news	117	102	5	1	1
editorial	12	12	1	1	1
forum	8	8	1	1	1
bio	5	4	2	1	1
faq	10	7	4	1	1
review	9	9	1	1	1
other	438	333	31	1	1

Max, min and med are abbreviations of minimum, maximum and median

from a single site (Wikipedia). The frequent inclusion of Wikipedia is most likely due to the popularity bias of current search engines.

In order to investigate the topic diversity of the LWGC-R corpus, we employed the technique described in Sect. 6.6 to extract keywords for the genre categories that comprise more than 1.5 % of the LWGC-R corpus. The results are presented in Table 19. Although the majority of the keywords are genre-specific, there are some topic-specific keywords such as "James LeBron" in the news articles. The reason for the presence of such topical keywords could be the recency bias of the collection method via search engines, i.e. collection at a particular point in time does not achieve temporal diversity. In future work, temporal diversity is therefore an additional factor that should be taken into account when collating a random genre corpus, i.e. the corpus collection should be performed at several time points instead of a single time point, at least for genres with a strong temporal connection such as news.

7.5 LWGC-R: extending coverage

Table 17 shows that 45.34 % of pages in LWGC-R did not belong to any of our 15 predefined genre categories, indicating a somewhat more than 50 % coverage for our 15 genres. 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 listed in Dimter (1981). Therefore, the web pages categorized as *other* in this experiment could belong to any genre class in these taxonomies.

New genre labels. In order to increase the coverage of genre annotation in the LWGC-R corpus, we investigated what genre classes the web pages annotated as

Table 19 Keywords from genre categories in LWGC-R which comprise more than 1.5 % of the corpus

blog (16)		com (167)		how-to (76)	
5	february	33	green	17	how
4	reply	20	services	15	writing
4	november	18	access	13	rules
4	do	16	statement	13	if
4	book	16	products	12	online
3	september	14	bank	12	freelance
3	posted	11	recycled	11	game
3	pm	11	leed	11	do
3	october	10	product	11	charge
3	news	9	*steel*	10	tips
3	march	9	recycling	9	article
3	lot	9	commitment	8	writer
3	january	9	banking	8	games
3	is	8	water	8	cards
3	if	8	systems	7	yourself
3	december	8	materials	7	writers
3	comment	8	business	7	make
3	by	7	support	7	learn
3	blog	7	manufacturing	7	job
3	big	7	estate	7	get
news (117)		**shop (79)**		**other (438)**	
41	news	18	products	43	are
24	said	18	product	35	edit
19	*james*	17	union	31	wikipedia
15	season	17	lack	31	blood
14	comments	16	accessories	30	sea
13	sports	15	see	29	can
11	*team*	15	price	28	was
10	*nba*	15	*clothing*	28	average
10	*lebron*	13	shop	27	be
10	*league*	13	shipping	26	environment
9	new	12	*shoes*	26	dictionary
9	cavaliers	11	customer	26	charge
8	wade	11	buy	26	by
8	state	10	supplies	25	as
8	reuters	10	star	24	do
8	points	10	mugs	23	will
8	percent	10	item	23	business

Table 19 continued

news (117)		shop (79)		other (438)	
8	national	10	amazon	22	what
8	mvp	9	reviews	22	this

Topic-specific keywords are indicated in italics

other mainly belong to. We observed that the class *other* consists of a considerable number of Wikipedia web pages and dictionary entries as well as directory web pages containing lists of links.[13] In addition, we could easily identify two genre categories *song lyrics* and *quotes*.

We tried to define these five genre classes as precisely as possible (Table 20). Then, we conducted another annotation experiment on MTurk in order to investigate how reliably humans can identify these additional five genre categories. The annotation procedure was exactly the same as the one described in Sect. 6.2 but was conducted only on the 438 pages in the LWGC-R gold standard previously defined as *other*.

Annotation results for new genre labels. For this experiment, the percentage agreement on 438 pages is 79.4 % and κ is 0.650 which indicates substantial agreement between the annotators (see Table 6).[14] Table 21, which depicts inter-coder agreement for the five individual categories, provides a more detailed picture of how reliable each genre class is.

While κ values for *quote*, *lyric* and *dictionary* are very high, and the value for *link lists* is substantial, the *encyclopedic articles* are not easy to identify reliably. Although naturally Wikipedia articles were easily identified as encyclopedic, there remained confusion between the border of an encyclopedic article and other informational descriptions as well as scientific articles. Figure 2 illustrates an example web page that creates such disagreement.

Table 22 shows the number of web pages for each of these five additional genre classes where at least three out of five annotations agreed. Adding these five genre classes to LWGC-R increases the genre coverage in this corpus to 74 %. Therefore, it is possible to extend the genre annotation coverage substantially.

Overall, the results show that our methodology of annotation can be expanded to more genre categories, although there are some genre classes that might not be suitable for MTurk annotation or need more clarification and refinement in terms of definition. It might also help to offer contrasting genre categories when introducing related genres (such as offering scientific articles as a contrast to encyclopaedic articles).

[13] Note that the inclusion of link lists departs from our original decision to focus on pages with large amounts of text. However, they seemed to be so frequent and popular that their inclusion might be necessary to enhance coverage.

[14] If we merge this new annotation with the previously conducted annotation on all 1000 web pages, overall κ using 20 categories and 1000 web pages is 0.67.

Table 20 The definition of additional genre classes

Genre	Definition
Dictionary/thesaurus entries	Explanations of a word's meaning and/or word translations and/or similar words. Includes pages where explanations in several dictionaries are listed
Link lists or directories of links	A page which consists mainly of links to other pages, which might be grouped topically or by genre (links to software downloads, for example). The start of the linked articles or documents might be included but not the full article. Tables of contents (if containing links) or indices (if containing links) are included
Song lyrics	The lyrics of one or more songs (not just links to such song lyrics)
Quotes	Including a single quote or a series of quotes
Encyclopedic articles	Contain an objective, non-opinionated description of entities such as (concrete and abstract) objects, organizations, places, events and animals. Most of the time, one of the first few sentences of these articles contains an objective definition of the entity described. Evaluative language in the definition such as "X is a must-have app" is not appropriate for an encyclopedia-like article. Although Wikipedia article pages are typical examples of this category, they are not the only encyclopedia-style articles on the Web. In fact, such descriptions do not even have to be published in an official encyclopedia. Some articles that are factual but do not qualify as encyclopedia-like articles are dictionary entries, announcements, or Wikipedia disambiguation pages

8 Conclusions and future work

In this paper, we present the first demonstrably reliably annotated web genre corpus. We developed precise and consistent annotation guidelines for well-defined and well-recognized categories. For annotating the corpus, we used crowd-sourcing. This avoids several problems in prior work such as annotation expense and speed. It also reduces dependency on experts and the resulting uncertainty about transferability of the annotation scheme to groups outside the development group.

Our corpus consists of two sub-corpora, of which one is created via focused search and the other via a more random sample of web pages returned by a search engine. Both are reliably annotated, showing that our annotation scheme is applicable to a wide range of arbitrary web pages as well. Both also are stored without information loss in HTML and visual format. The focused search sub-corpus has a reasonable number of pages for each genre category which is important for training machine learning algorithms. Both corpora are source and topic-diverse, although the random sub-corpus has limited temporal diversity, leading to lack of topic diversity for a single genre (news), which should be addressed in future extensions. We have also shown that our annotation approach can be extended to include further genre categories and therefore extend genre coverage. However, great care needs to be taken to offer very precise category definitions for naive annotators, and each new genre category needs to be checked for reliability.

Table 21 Inter-coder agreement for the additional genre classes in LWGC-R

Genre labels	Percentage agreement	Fleiss's κ
Encyclopedia-type articles	68.8	0.582
Dictionary/thesaurus entries	96.1	0.767
Link lists or directories of links	87.3	0.658
Song lyrics	99.3	0.733
Quotes or lists of quotes	98.6	0.873
Other	64.4	0.639

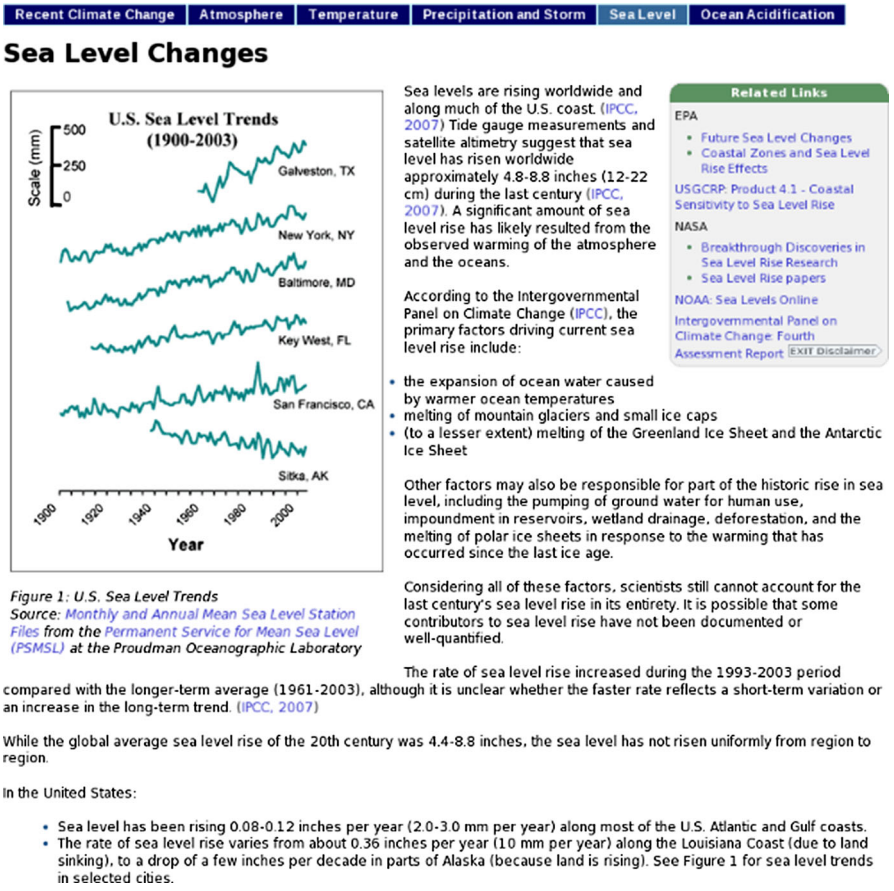

Fig. 2 An example web page which causes confusion between the classes *Encyclopedia-type articles* and *other*. http://epa.gov/climatechange/science/recentslc.html

An important future direction lies in expanding the corpus. Increasing the amount of data can be beneficial for machine learning algorithms (Banko and Brill 2001). Therefore, we should expand the corpus in terms of size which could be done via

Table 22 Distribution of the additional genre classes in the LWGC-R

Category	# Web pages	% of the corpus
Encyclopedia-type articles	97	9.7
Link lists or directories of links	56	5.6
Dictionary/thesaurus entries	22	2.2
Quotes or lists of quotes	13	1.3
Song lyrics	3	0.3
Other	241	24.1

focused search (as for LWGC-B) or by annotating random web pages (as for LWGC-R). Both of these approaches have advantages and disadvantages. While extending the corpus using random web pages results in an unbalanced corpus, it eliminates expert selection bias by the development group and includes less prototypical examples of genre categories. On the other hand, by employing a focused-search approach, we can create a balanced corpus and overcome the problems that a skewed corpus can create for machine learning algorithms. Therefore, we think extending the corpus should be done by employing both of these approaches. We also think that in addition to source and topic diversity, other variables should also be controlled, such as temporal diversity.

Another way of extending the corpus is to increase the number of genre categories. We show that our original 15 genre categories are sufficient to cover the majority but not the vast majority of web pages and our extended inventory of 20 genre categories covers about three quarters of web pages. As noted in Sect. 5, there is no universally agreed set of genre labels. However, as long as the web users can identify a genre category reliably in an annotation task, it can be added to the corpus. When extending the genre categories, the issues of granularity and a potential hierarchical organisation will need to be investigated.

One other issue of corpus extension is to create a multilingual genre corpus. Currently, we only concentrated on English web pages. It would be interesting to see how genres differ cross-culturally.

Acknowledgments This research was partly funded by a Google Research Award to Serge Sharoff and Katja Markert. Noushin Rezapour was funded by an EPSRC Doctoral Training Grant.

References

Andersen, J. (2008). The concept of genre in information studies. *Annual Review of Information Science and Technology*, *42*(1), 339–367.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pp. 26–33. Association for Computational Linguistics

Baroni, M., & Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. *Proceedings of LREC*, *4*, 1313–1316.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Beigman Klebanov, B., & Beigman, E. (2009). From annotator agreement to noise models. *Computational Linguistics*, *35*(4), 495–503.

Bennett, E., Alpert, R., & Goldstein, A. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, *18*(3), 303.

Bentivogli, L., Federico, M., Moretti, G., & Paul, M. (2011). Getting expert quality from the crowd for machine translation evaluation. *Proceedings of the MT Summmit*, *13*, 521–528.

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., & Secret, A. (1994). The world-wide web. *Communications of the ACM*, *37*(8), 76–82.

Berninger, V., Kim, Y., & Ross, S.(2008). Building a document genre corpus: A profile of the KRYS I corpus. In *Proceedings of the BCS-IRSG workshop on corpus profiling*.

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Callison-Burch, C. (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, pp. 286–295. Association for Computational Linguistics.

Campbell, K. K., & Jamieson, K. H.(1978). Form and genre in rhetorical criticism: An introduction. Form and genre: Shaping rhetorical action pp. 9–32.

Cohen, J., et al. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Crowston, K., Kwaśnik, B., & Rubleske, J. (2011). Problems in the use-centered development of a taxonomy of web genres. Genres on the Web pp. 69–84.

Crowston, K., & Kwasnik, B. H.(2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th annual Hawaii international conference on system sciences, IEEE.*

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the world wide web. *The Information Society*, *16*(3), 201–215.

Davies, M., & Fleiss, J. (1982). *Measuring agreement for multinomial data.* Biometrics, pp. 1047–1051.

Dillon, A., & Gushrowski, B. A. (2000). Genres and the web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science*, *51*(2), 202–205.

Dimter, M. (1981). Textklassenkonzepte heutiger Alltagssprache: Kommunikationssituation, Textfunktion und Textinhalt als Kategorien alltagssprachlicher Textklassifikation, vol. 32. Walter de Gruyter.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Egbert, J., & Biber, D. (2013). Developing a user-based method of register classification. In *Proceedings of the 8th web as corpus workshop*. Lancaster.

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*.

Meyer zu Eissen, S., & Stein, B. (2004). Genre classification of web pages. KI 2004: Advances in Artificial Intelligence pp. 256–269.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370. Association for Computational Linguistics.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Giesbrecht, E., & Evert, S. (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)*, pp. 27–36.

Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, *32*(1), 83–135.

Kaisser, M., Hearst, M., & Lowe, J. (2008). Evidence for varying search results summary lengths. In *Proceedings of ACL*, pp. 701–709.

Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing & Management*, *45*(5), 499–512.

Kessler, B., Numberg, G., & Schutze, H. (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*, pp. 32–38. Association for Computational Linguistics.

Kilgarriff, A. (2012). Getting to know your corpus. In *Text, speech and dialogue*, (pp. 3–15). Springer.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology5*(3), 37–72. http://llt.msu.edu/vol5num3/pdf/lee.pdf.

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, *41*(5), 1263–1276.

Litman, D., Hirschberg, J., & Swerts, M. (2006). Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, *32*(3), 417–438.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods*, *44*(1), 1–23.

McCreadie, R., Macdonald, C., & Ounis, I. (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)*, pp. 23–26.

Mehler, A., Sharoff, S., & Santini, M. (Eds.). (2010). *Genres on the web: Computational models and empirical studies*. Berlin, New York: Springer.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, *70*(2), 151–167.

Mohammad, S., & Turney, P. (2012). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Orlikowski, W., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. Administrative science quarterly. pp. 541–574.

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics*, *37*(2), 385–393.

Rayson, P., Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora*, pp. 1–6. Association for Computational Linguistics.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., & Vidulin, V. (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, pp. 351–358.

Riezler, S. (2014). On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, *40*(1), 235–245.

Santini, M.(2007). Automatic identification of genre in web pages. Ph.D. thesis, University of Brighton.

Santini, M. (2008). Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing & Management*, *44*(2), 702–737.

Santini, M., Mehler, A., & Sharoff, S. (2010). Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the web: Computational models and empirical studies*. Berlin, New York: Springer.

Scott, W.(1955). Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. WaCky, pp. 63–98.

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and internet. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the web: Computational models and empirical studies* (pp. 149–166). Berlin, New York: Springer.

Sharoff, S., Wu, Z., & Markert, K. (2010). The web library of babel: evaluating genre collections. In *Proceedings of the seventh conference on international language resources and evaluation*, pp. 3063–3070.

Shepherd, M., & Watters, C.(1998). The evolution of cybergenres. In *System Sciences, 1998, Proceedings of the thirty-first Hawaii international conference*, vol. 2, pp. 97–109. IEEE.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (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*, pp. 254–263. Association for Computational Linguistics.

Snyder, B., & Palmer, M. (2004). The english all-words task. In *Senseval-3: Third international workshop on the evaluation of systems for the semantic analysis of text*, pp. 41–43.

Steger, J. M., & Stemle, E. W. (2009). KrdWrd - architecture for unified processing of web content. In *Proceedings of the fifth web as corpus workshop*, pp. 63–70.

Stubbe, A., & Ringlstetter, C.(2007). Recognizing genres. In *Proceedings of the colloquium towards a reference corpus of web genres*.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Vidulin, V., Luštrek, M., & Gams, M. (2007). Using genres to improve search engines. In *Towards genre-enabled search engines: The impact of natural language processing*, pp. 45–51.

Vieira, R., & Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, *26*(4), 539–593.

Webber, B.(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, pp. 674–682. Association for Computational Linguistics.

Yates, J., Orlikowski, W., & Rennecker, J. (1997). Collaborative genres for collaboration: Genre systems in digital media. In *System sciences, 1997, Proceedings of the thirtieth Hawaii international conference*, vol. 6, pp. 50–59, IEEE.