**Article:**
Sharoff, S (2018) Functional Text Dimensions for the annotation of web corpora. Corpora, 13 (1). pp. 65-95. ISSN 1749-5032

https://doi.org/10.3366/cor.2018.0136

# Functional Text Dimensions for annotation of Web corpora*

Serge Sharoff
University of Leeds
s.sharoff@leeds.ac.uk

**Abstract**

This paper presents an approach to classify large Web corpora into genres by means of Functional Text Dimensions (FTDs). This offers a topological approach to text typology in which the texts are described in terms of their similarity to prototype genres. The suggested set of categories is designed to be applicable to any text on the Web and to be reliable in annotation practice. Interannotator agreement results show that the suggested categories produce Krippendorff's $\alpha$ above $0.76$. In addition to the functional space of 18 dimensions, similarity between annotated documents can be described visually within a space of reduced dimensions obtained through t-distributed Statistical Neighbour Embedding. Reliably annotated texts also provide the basis for automatic genre classification, which can be done in each FTD, as well as as within the space of reduced dimensions. An example comparing texts from the Brown Corpus, the BNC and ukWac, a large Web corpus, is provided.

## 1 Introduction

Many corpora collected from the Web, like ukWac (Baroni et al., 2009), lack even the most basic information about their genres, e.g., the proportion of news items, academic writing or fiction. Given that many linguistic features vary according to genres, for example, this concerns the word frequencies, as well as the use of pronouns or passive voice (Biber, 1988), the utility of Web corpora can be enhanced by understanding their composition, selecting their genre-specific subsets, or comparing one Web corpus against the other in terms of the genres they contain.

However, getting a suitable set of genre labels for a Web corpus is surprisingly difficult. The text users easily navigate in a wide range of genres of everyday life, for example, emails, newspaper columns, feature articles, newswires, advertising or instruction manuals. The names given to these genres by the users provide the primary evidence for their existence as genres (Waller, 2012). However, the users do not have operational definitions of what these names mean precisely. Also each user usually concentrates on a small number of types of texts relevant to *their* everyday life, providing

*Accepted for publication in the *Corpora* journal, http://www.euppublishing.com/loi/cor. Please cite this paper as 'Sharoff, S. (2018). Functional Text Dimensions for annotation of Web corpora. *Corpora*, **31**:2.'

situation-specific labels, such as 'uncontrolled resource page' or ambiguous ones, such as 'article', thus necessitating more research into linking the genre labels to the way they are actually used. For more information on the problems with the user-based genre taxonomies see (Crowston et al., 2010).

At the same time, exhaustive genre inventories covering all possible genres tend to develop into unstructured alphabetically ordered lists, such as those used in traditional studies of text typology, for example, more than 4,000 items in the list of (Adamzik, 1995), which includes such genres as *Abschiedsbrief* 'suicide note', *Abschiedsgespräch* 'farewell talk' or *Abschiedsrede* 'farewell address'. A slightly shorter list of about 2,000 genre labels in English is analysed in (Görlach, 2004).

The task of comprehensive coverage for a general-purpose corpus necessitates a longer list of labels, while a small number of labels is needed for selecting reasonably sized subcorpora and comparing language use between them. Besides the fact that genre inventories containing thousands of labels are too long for practical purposes of text classification, data from the Web can present a wider variety of types of texts than available even in those inventories, for example, blogs, emails and Wikipedia articles are not listed in (Adamzik, 1995) for obvious reasons. Another difficulty of dealing with long lists of genre labels is a high degree of genre hybridism, especially on the Web, where many texts are not controlled by the institutional gate-keepers and can appear in many forms, thus possibly violating well-defined genre conventions or blending them (Santini et al., 2010).

Traditional corpora, like the Brown Corpus (Kučera and Francis, 1967), or the BNC (Lee, 2001), have been provided with their own classification schemes. However, their sets of genre categories differ without even minimal compatibility: 15 categories are provided in the Brown Corpus, including six categories of fiction, vs 70 categories in the BNC with only one category for fiction. In comparison to these relatively small pre-designed collections, large corpora collected as a snapshot of Web texts or by exhaustive crawling need a larger set of labels to cover the majority of their texts. Even in the BNC genre index (Lee, 2001), the most common genre category is the non-informative `W_misc` (501 texts).

One of the important lessons identified in the process of annotation experiments concerns the reliability of assigning a label (Artstein and Poesio, 2008), a problem, which is especially relevant in the case of labelling texts by their genres. If a webpage naturally allows two interpretations, two annotators or even the same annotator annotating the same text for the second time can reasonably consider two different labels. For example, the three categories of Newspaper texts defined in the Brown corpus are: **A** Reportage, **B** Editorials, and **C** Reviews, the boundaries between which are not always clear. In addition to having purely reporting texts in Category A, some texts can be reasonably considered as belonging to Category B, such as Text `A`04:

> *The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer, franker, and wider discussions, animated by much better mutual understanding than in past meetings. This has been a working session of an organization that, by its very nature, can only proceed along its route step by step and without dramatic changes...* ("NATO Welds Unity" from The Christian Science Monitor)

Any discrepancy in annotation of such pages makes it harder to assess the composition of a corpus and to train text classification tools in the Machine Learning framework. For example, if two annotators do not reach agreement on what is considered to be a news report from what is considered to be an opinion piece for an arbitrary newspaper text, it is impossible to make any claim about the statistical significance of differences in their linguistic features.

While the web users can talk confidently about the genres of texts they deal with in their daily lives, they do not necessarily agree when they apply labels to classify them. On randomly selected webpages the level of interannotator agreement in existing genre collections with multiple annotations is usually below the commonly accepted reliability thresholds (Sharoff et al., 2010).

Finally, traditional corpora provide no estimation of how similar different texts within the same categories are. For example, Category J ('learned') in the Brown Corpus contains research articles in radio engineering, chemistry and psychoanalysis, as well as essays on history, opera and poetry. Arguably they are more dissimilar within this category in comparison to samples from two separate categories L (Mystery&crime fiction) and N (Adventure fiction). Similarly, the third most common category in the BNC is `W_pop_lore` (211 texts), which codes a variety of news outlets, like *The Economist, She Magazine, Amnesty International* or *Climber and Hill Walker*. While genre-wise they are quite different from each other, even within each of these magazines we can find a variety of genres. For example, the texts from the *Climber and Hill Walker* magazine contain news items, equipment advice, personal stories or invitations to joining a campaign. Ideally, different corpora also need to be compared with respect to their genre composition using a shared set of genre categories, for example, for comparing ukWac to the BNC or the Brown Corpus. For example, one can expect similarities between argumentative blogs from webcorpora and argumentative texts in traditional corpora, for example, those in Category B of the Brown Corpus (Press/Editorials).

In spite of all these difficulties, the intuition for corpus collection assumes a genre classification framework: texts naturally belong to different genres, and the difference between their genres is reflected in different linguistic features used. Therefore, a corpus is expected to have an explicit representation of the types of texts it contains. Traditional general-purpose corpora usually have a genre typology. Even in specialised corpora like the Wall Street Journal corpus, which look homogeneous, one can find such different genres as opinion pieces, financial reports or letters to the editor, which, among other things, exhibit very different inventories of discourse relations (Webber, 2009).

Unlike traditional corpora, Web-derived corpora usually do not have labels apart from their provenance from a particular website. While the base URL address of a webpage hints at some genre restrictions, for example, it is less likely to find recipes in a corpus coming from a university website,[1] many other genres are possible in an academic corpus, such as biographies, project descriptions, legal notices, technical support, news messages, obituaries or blog entries. A blog entry in itself is not a definite genre either. Texts published in a blog can be personal diaries, polemic discussions, instructions, messages for information or advertising, invitations for participation, etc.

[1]However, `http://www.leeds.ac.uk/youarewhatyouate/Recipes/` demonstrates that even this unlikely genre can be found in a corpus of academic webpages.

John Sinclair introduced a useful distinction between text-external vs text-internal criteria for text classification (Sinclair and Ball, 1996). A typical text-external classification parameter concerns properties of the author, e.g., the age or the place of birth. A typical text-internal parameter is classification of topics by keywords extracted from the text. If we apply this distinction to genres, they can be defined both externally and internally. Text-external criteria for defining genres refer to the communicative aims as intended by the author or perceived by the audience. Text-internal criteria refer to the lexicogrammatical choices made within texts, such as the use of contractions, past tense verbs or clause coordination. Traditional genre labels are mostly defined text-externally, see (Görlach, 2004), while many linguistic approaches to genre analysis investigate genres text-internally (Biber, 1988). Cf also a discussion of the difference in uses of the terms 'genre' (text-external) vs 'register' (text-internal) in (Lee, 2001).

The study presented here treats the two perspectives as complementary. It starts from the text-external end, so that the genres are defined via generalised communicative aims and these definitions avoid references to their lexicogrammatical realisations. This is also motivated by the need of doing genre analysis across languages, so that a collection of English webpages can be compared to two collections of Chinese and Russian ones. Once a reliable text-external framework has been established, we will be able to investigate its correlation with any appropriate text-internal features in a given language, in particular with the aim of developing an automatic classification tool.

This paper introduces the framework of Functional Text Dimensions (FTDs, Section 2), presents argumentation concerning the choices made in developing the FTD set used in this study in comparison to other genre studies (Section 3), reports the results of an interannotator agreement study for the FTDs (Section 4), presents analysis of FTD clusters, which correspond to established Web genres (Section 5), and also investigates ways for automatic genre classification using Machine Learning (Section 6).

## 2 Functional Text Dimensions

The aim of this study is to develop a text-external genre classification framework for Web corpora, which satisfies the following requirements:

- it is applicable to the majority of modern texts on the Web;
- it can describe any text in a sensible way using a relatively small number of parameters;
- it ensures that several annotators with sufficient training consistently produce the same annotation for the same text.

The first requirement means that the scheme is aimed at describing the composition of large corpora of general language. The annotation scheme should produce useful codes to the majority of their texts, while specialised corpora might need a more fine-grained scheme. The second requirement is aimed at selecting and comparing subsets of Web corpora, which is more difficult to achieve with large genre inventories. Having fewer categories should also simplify the manual annotation process. This is also likely to lead to improved automatic classification results in the Machine Learning framework. The third requirement is aimed at greater confidence in linking human perception of genres to the linguistic features in texts. This also helps in automatic classification.

This paper describes Functional Text Dimensions (FTDs) as one of the ways for satisfying the diverse requirements for genre classification. First, the FTDs offer a compact set of relatively general functional categories instead of atomic labels, so that the genre of a text can be described through a combination of several parameters, as this is done in componential analysis in phonetics and semantics. Decomposition of labels reduces the number of descriptive parameters in comparison to atomic labels. It also helps in comparing text collections across different genre frameworks, if their labels can be decomposed in compatible ways.

Second, the notion of distance is introduced: the value of an FTD for each text is measured on a scale of how strong this text features in this dimensions. Instead of a typology, which describes a text as a member of genre classes, FTDs describe a text topologically in the space of functional dimensions, so that we can measure its distance from what can be considered as prototypical texts.

Examples of FTDs can be 'informative reporting' or 'argumentation'. A typical newswire features strongly on the 'informative reporting' dimension, while it gets 0 on the 'argumentation' dimension. Some texts, for example, Text `A04` from the Brown Corpus discussed above, get a non-zero score on both dimensions, thus removing the need of labelling it as either Category A (Reportage) or B (Editorial). This allows straightforward representation of genre hybridisation, which is especially common in texts on the Web.

In annotation practice, the FTDs are operationalised via test questions, for example:

**A8: hardnews** To what extent does the text appear to be an informative report of events recent at the time of writing? (For example, a news item).

**A1: argum** To what extent does the text contain explicit argumentation to persuade the reader? (For example, argumentative blogs, opinion pieces or discussion forums).

**A17: eval** To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, a product review)

Each FTD is best described via its test question. For presentational reasons, the FTDs also have codes (**A1, A8**, etc), as well as short labels (`argum, hardnews`, etc). The test question also indicates the prototypical genres for an FTD, so that the value a text gets on this FTD is judged with respect to its similarity to the prototypical genre listed after the test question. All of the FTDs taken together provide a multi-dimensional representation for positioning any text as a point in this space. If two texts are perceived as similar in terms of their genres, they are likely to be quite close to each other in the FTD space. Unlike the traditional approach to designing genre lists or hierarchies in which texts need to be members of the respective classes, the FTD approach to detecting the genre of annotated texts is based on their similarity to prototypes, which is measured as distance in the FTD space.

Analysis of text genres as presented in this paper is based on four FTD groups, which offer a way of organising more specific FTDs and their prototypical genres:[2]

[2]The test questions for the FTDs are given in Appendix 1.

**information** General functional categories aimed at informing the reader:

**A7 instruct** Tutorials, FAQs, manuals, recipes;

**A8 news** Newswires, newsletters;

**A9 legal** Laws, contracts, small print;

**A16 info** Specifications, CVs, encyclopedic articles, abstracts;

**discussion** General functional categories aimed at discussing a state of affairs:

**A1 argum** Editorials, columns, argumentative blogs, political debates;

**A14 research** Research articles, essays;

**A17 eval** Reviews of products and services;

**narration** General functional categories aimed at presenting a story

**A4 fiction** Novels, stories, verses;

**A11 person** Diary-like blogs, personal letters, traditional diaries;

**promotion** General functional categories aimed at promotion of information:

**A12 commpuff** Advertising and commercial texts for promotion;

**A13 ideopuff** Propaganda, manifestos;

**A20 appell** Requests, small ads.

These twelve FTDs can be treated as the *principal* dimensions: a text needs to be positioned at a non-zero point along at least *one* of them. In addition to them, there can be important variation within texts along some other dimensions, including:

**A3 emotive** Texts expressing feelings and emotions;

**A5 flippant** Light-hearted texts aimed at entertaining the reader;

**A6 informal** Communication arising from a relaxed situation, deviating from the accepted "prestige" standard;

**A15 specialist** Texts requiring background knowledge for their understanding;

**A18 dialogue** Texts containing active interaction between several participants.

The specific set of dimensions listed here came from the results of several annotation experiments, primarily (Sharoff, 2010; Forsyth and Sharoff, 2014; Sorokin et al., 2014; Katinskaya and Sharoff, 2015). Each annotated text in those experiments has been positioned along at least one of the principal dimensions, which provide the primary reference point for describing text functions. In some cases, two texts having the same function can differ stylistically. The most common cases found in annotation are captured by the secondary dimensions, e.g., texts written with entertainment purposes (**A5**) or texts for specialists (**A15**).

Human annotators have been asked to evaluate texts for each of these dimensions by answering the test questions on a customised Likert scale:

**0** none or hardly at all;

**0.5** slightly;

**1** somewhat or partly;

**2** strongly or very much so.

The default value for an FTD is 0 ('None'). When a text is to some extent similar to a prototypical genre, the corresponding FTD can be set to an appropriate value. The annotation scale is designed to force the annotator to declare the presence of a functional dimension in a text and assign 'Strongly' if this is the case. The intention was to emphasise the difference in the presence of a particular dimension in a text: the text is either a recognised representative of a relevant genre ('I know it is this genre when I see it'),[3] or it is related to this genre to a lesser extent, in the latter case the value is 'Partly'. The greater numerical difference between 'Strongly' (2) and other values on the scale is intended to help with providing better separation between the resulting genres in the multidimensional space. As for the number of items, research into questionnaire design suggests that "scale-level scores are possibly being influenced by 'untrue' middle response category endorsement," when the respondents treat it as a "dumping ground" (Kulas et al., 2008). Therefore, the genre assessment scale for this study was designed to include an even number of points.

While the FTD descriptions are similar to genre labels, the FTDs offer the possibility to generalise some of the traditional labels, such as 'Persuasive article,' 'Editorial' and 'Opinion blog' (Egbert et al., 2015) into **A1**, while also finding variations between them to distinguish between purely informational presentation of news[4] vs a combination of informational presentation with argumentation[5] by setting the appropriate values of **A8**.

As an example of annotation, consider a blog entry which begins with:[6]

> There were quite a few Kexi releases since my last blog entry. I tell you, the focus in this work was on improving stability. As an effect, reportedly, there can be a whole day of work without stability issues. Not bad.. Kexi is in fact a family of 5 or more apps integrated into ...

It can receive such annotations as:
`inform: 2, instruct:1, person:1, appell:1, specialist:0.5, flippant:0.5`

The text clearly features on **A16** (information). In addition to this, it can be considered as having a non-zero value in three other dimensions, **A7** (instructions), **A11** (reporting personal experience) and **A20**, an appeal to support an open-source project. In a more traditional annotation scheme with atomic labels, e.g., the one used in a large-scale Web annotation study (Egbert et al., 2015), a text like this is likely to be annotated as an 'informational blog', which is acceptable, but this does not cover the

[3] Cf `https://en.wikipedia.org/wiki/I_know_it_when_I_see_it`

[4] For example, `http://en.wikinews.org/w/index.php?oldid=1108629`

[5] `http://www.project-syndicate.org/print/the--browning--of-african-technology`

[6] `https://blogs.kde.org/2014/09/04/kexi-gsoc-jj-porting`

more fine-grained annotations and does not relate it as partly similar to other category labels available in the same genre inventory, e.g., 'Technical support' and 'Instructions'.

Overall, a 'blog entry' or a 'Twitter message' is not a genre label, this is a way of content distribution, while its FTD is best described via the function of the message, e.g., argumentation (**A1**, for example, a political discussion), distribution of news (**A8**) or request (**A20**).

FTD-based annotation has been tested on a range of corpora, also covering several languages. The main test set has three components (see Table 1):

**5g** the Pentaglossal corpus, which consists of texts with their translations in five languages (Chinese, English, German, French and Russian); the collection contains texts coming from fiction, corporate communication, political debates, TED talks, UN reports, etc (Forsyth and Sharoff, 2014);

**ukWac** randomly selected texts from ukWac, an English-language webcorpus collected in 2006 by crawling the `.uk` domain (Baroni et al., 2009);

**GICR** annotated texts from GICR, a Russian-language webcorpus collected in 2013 (Piperski et al., 2013); some of the annotated texts have been selected from a targeted search to represent a wide variety of genres, while some have been randomly selected from the most popular internet sources, such as news, Wikipedia or several blogging platforms.

In terms of genre coverage, the current set of FTDs ensures that the texts from collections annotated in Table 1, as well as some texts from traditional corpora, such as BC and BNC, can receive an annotation along at least one principal dimension. Automatically generated texts, such as Web search results, are usually excluded from general-purpose corpora, so they have not been considered in annotation.

Table 2 presents an example of corpus composition analysis using the FTDs. Each row shows the overall sum of all annotations for an individual dimension in each corpus, as well as the mean value in each dimension for this corpus. This can help in comparing corpora in terms of their most significant FTDs. Since the FTDs and their annotation guidelines refer to text-external parameters, it is possible to compare the composition of corpora across languages, especially when they are nearly contemporaneous text collections produced using similar pipelines, like ukWac and GICR. The three most typical FTDs in the annotated subset of ukWac are news (**A8**), commercial promotion (**A12**) and information (**A16**). The random portion of annotated GICR texts is broadly similar to ukWac. The differences concern a greater amount of informative texts, mostly from Wikipedia, which is missing from ukWac, since these texts are not coming from the `uk` domain. GICR (random) also has more argumentative texts (**A1**) in comparison to news, as well as diary-like reporting (**A11**). At the same time, the blog portion of GICR is markedly different by having more personal reporting and argumentative texts, often mixed with reviews of various kinds.

## 3 Related studies

The 'jungle' metaphor is quite common in the field of text classification (Sharoff, 2010). A general overview of the approaches and variation in terminology used is available in

| Corpus | Language | #Docs | #Words |
|---|---|---|---|
| 5g, part1 | en,de,fr,ru,zh | 113 | 306302 |
| 5g, part 2 | en,fr,ru,zh | 133 | 505468 |
| ukWac, random | en | 257 | 211549 |
| GICR, random | ru | 618 | 919972 |
| GICR, blogs | ru | 285 | 83829 |
| Total | | 1406 | 2027120 |

Table 1: Corpora used in this study

| FTDs | A1 | A4 | A7 | A8 | A9 | A11 | A12 | A13 | A14 | A16 | A17 | A20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ukWac, rand | 55.5 | 9.0 | 79.5 | 102.5 | 44.0 | 32.0 | 125.0 | 21.0 | 31.0 | 136.0 | 55.0 | 44.0 |
| mean | 0.22 | 0.04 | 0.31 | **0.40** | 0.17 | 0.12 | **0.49** | 0.08 | 0.12 | **0.53** | 0.21 | 0.17 |
| GICR, rand | 261.0 | 59.0 | 77.0 | 190.0 | 157.0 | 238.0 | 138.5 | 134.5 | 124.0 | 549.0 | 230.0 | 56.0 |
| mean | **0.42** | 0.10 | 0.12 | 0.31 | 0.25 | **0.39** | 0.22 | 0.22 | 0.20 | **0.89** | 0.37 | 0.09 |
| GICR, blogs | 108.5 | 54.5 | 46.0 | 23.0 | 8.5 | 345.5 | 51.5 | 39.0 | 15.0 | 40.0 | 241.0 | 12.0 |
| mean | **0.38** | 0.19 | 0.16 | 0.08 | 0.03 | **1.21** | 0.18 | 0.14 | 0.05 | 0.14 | **0.85** | 0.04 |

Table 2: Comparing corpus composition in terms of FTDs

(Lee, 2001) and in (Biber and Conrad, 2009). Even though there is no consistency in the use of terminology between genre researchers, the discussion below tries to maintain consistency by referring to genres (or FTDs) as text-external parameters, while registers are considered to be defined text-internally. This differs from the text-external definition of registers in (Biber, 1988), but keeps compatibility with the discussion of genres and registers in (Halliday, 1985; Lee, 2001; Santini et al., 2010).

The FTD list introduced above continues a line of genre classification experiments, which started from adaptation of John Sinclair's typology of communicative aims (Sinclair and Ball, 1996) to the needs of the Russian National Corpus (Sharoff, 2005). Among other parameters of text classification Sinclair referred to the following six 'intended outcomes of text production':

**information** reference compendia (Sinclair adds the following comment "an unlikely outcome, because texts are very rarely created merely for this purpose");

**discussion** polemic, position statements, argument;

**recommendation** reports, advice, legal and regulatory documents;

**recreation** fiction and non-fiction (biography, autobiography, etc)

**religion** holy books, prayer books, Order of Service (this label does not refer to religion as a topic);

**instruction** academic works, textbooks, practical books.

The typology is compact: it contains only six top-level categories, each of which can be used to describe a variety of webpages, e.g., a page from Wikipedia is aimed

at informing, a forum — at discussing, etc. However, some webpages do not receive a suitable category from this list, for example, adverts, while the annotations of other webpages can be ambiguous, for example, reports vary in the degree of information, discussion or recommendation they contain.

Experience in genre annotation of several general-purpose corpora collected from the Web (Sharoff, 2006) led to the following Functional Genre Classes (Sharoff, 2010):

**information** (catalogues, glossaries, home pages)

**discussion** with subcategories:
- **academic** (research papers and monographs)
- **public** (journalism and political debates)
- **everyday communication** (forums, emails, diary blogs)

**reporting** (newswires, police reports)

**instruction** (how-tos, FAQs, tutorials)

**regulations** (laws, small print, contracts)

**promotion** (adverts, political propaganda)

**recreation** (fiction and popular lore)

**non-text** (pages with little running text):
- applications (Flash, Java, applets); online interfaces (forms for making queries, purchases, downloading, or logging in); linkerie (portals, link lists)

This typology has been designed specifically to cover a wide range of typical webpages. Various samples of webpages totalling in 2,000 have been classified using this scheme, which was shown to be reasonably useful, but not reliable with respect to inter-annotator agreement. While double annotation of 200 random webpages using FGC leads to agreement within the acceptable threshold of Krippendorff's $\alpha = 0.70$, only the *regulations* category is reliable (.93); other categories result in considerably lower inter-annotator agreement figures (Sharoff et al., 2010):
*reporting* (.57), *discussion* (.64) and *promotion* (.55)

The lack of reliability in classification of random web texts as well as the need to account for variation within each genre category led to a proposal for introducing FTDs. The initial set of 17 dimensions as initially proposed in (Forsyth and Sharoff, 2014) has been revised in several annotation experiments (Sorokin et al., 2014). Some dimensions exhibited considerable correlation with each other, so they were considered to be redundant, while some texts did not receive a principal dimension in the preliminary experiments. This led to adding more dimensions. The current FTD set with their definitions is listed in Appendix 1 in this paper.

Two approaches to genres are particularly relevant to the proposed framework. One is the Multi-Dimensional Approach pioneered by Biber, e.g., (Biber, 1988; Biber, 1995; Grieve et al., 2010), which also involves judging types of texts along dimensions. MDA also proposes ways of accounting for variation, i.e., the similarities and differences in individual texts either between texts of different categories or within the same category.

The main difference of the proposed approach from MDA is that the FTDs are based on text-external definitions, i.e., on the communicative function of a text, while Biber's text dimensions are derived text-internally, i.e., from co-occurrences of linguistic features. The two views are complementary, and the study presented in Section 6 is based on text-internal parameters of genres. However, the text-internal dimensions discovered in MDA need to be validated through human perception of how different the texts are. This is why MDA studies start with text-external definitions of text categories, which are described by Biber as 'registers'. However, these categories are usually defined according to the sources of data, so they are not compatible across MDA studies. The contribution of the FTDs consists in decomposition of the MDA 'registers' to provide a more stable way of text-external descriptions across a range of text collections.

Another related approach concerns a proposal to treat text typology as a 'topology', which has been suggested in (Lemke, 1999). The semiotic topography discussed by Lemke addresses both the text-external dimensions by considering the similarity of texts with respect to their communicative intentions, and the text-internal dimensions by considering the similarity of their linguistic structures. Genetically this comes from the same theoretical background as (Halliday, 1985; Sinclair and Ball, 1996), since it links the text-internal choices, which depend on lexicogrammar, to the text-external choices, which depend on the context of use. However, the early proposal by Lemke did not define an explicit set of suitable dimensions for a general corpus other than discussing individual examples, such as stories.

# 4 Reliable corpus annotation

Given that unreliable annotation is a major issue in genre research, each genre framework needs to be tested for reliability. One way to measure interannotator agreement is by using Krippendorff's $\alpha$ (Krippendorff, 2004), which treats the annotators as interchangeable and measures the difference between disagreement expected by chance vs observed disagreement:

$$\alpha = 1 - \frac{D_{observed}}{D_{chance}} = 1 - \frac{\sigma^2_{within}}{\sigma^2_{total}}$$

where $\sigma_{within}$ is standard deviation of the differences within the annotations for the same text, $\sigma_{total}$ is standard deviation of the overall difference between all annotations. The level of Krippendorff's $\alpha$ of 1 indicates perfect agreement, i.e., there is no variation in assessing the same item between the annotators. According to Krippendorff, the threshold of $\alpha \geq 0.80$ indicates reliable judgements, while $\alpha \geq 0.67$ is recommended as a limit to support conclusions about reliability of annotation, because the value of $\alpha$ below this limit makes conclusions drawn from respective data not significant from the statistical viewpoint.

The reason for choosing Krippendorff's $\alpha$ over other reliability measures is because it naturally takes into account the **degree** of disagreement. For example, the difference in choosing 'None' (0) vs 'Strongly' (2) should be more important than the difference in choosing 'None' (0) vs 'Slightly' (0.5).

The annotators were MA students in Linguistics or in Translation Studies. They received initial training in using the FTDs, which consisted of:

| | **A1** | **A3** | **A4** | **A5** | **A6** | **A7** | **A8** | **A9** | **A11** | **A12** | **A13** | **A14** | **A15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5g, p1 | 0.91 | 0.79 | 0.97 | 0.69 | 0.69 | 0.98 | 0.86 | 0.89 | 0.80 | 0.88 | 0.93 | 0.90 | 0.59 |
| 5g, p2 | 0.80 | 0.95 | 1.00 | 0.97 | 0.91 | 0.99 | 0.78 | 1.00 | 0.94 | 1.00 | 0.89 | 0.91 | 0.90 |
| GICR, rand | 0.94 | 0.97 | 0.71 | 0.75 | 0.84 | 0.82 | 0.99 | 1.00 | 0.74 | 0.66 | 0.67 | 0.77 | 0.84 |
| GICR, blogs | 0.83 | 0.70 | 0.77 | 0.62 | 0.67 | 0.98 | 1.00 | 1.00 | 0.80 | 0.60 | 0.66 | 0.61 | 0.77 |
| **Mean** | **0.87** | **0.85** | **0.86** | **0.76** | **0.78** | **0.94** | **0.91** | **0.97** | **0.82** | **0.78** | **0.79** | **0.80** | **0.78** |

Table 3: Corpora with multiple annotations and the Krippendorff's $\alpha$ for FTDs

1. a presentation explaining the FTD test questions and their prototypical texts;
2. a joint annotation session, in which the FTD values for 10 texts were discussed;
3. an independent annotation of 25 texts selected to represent a variety of typical texts on the Web;
4. a norming session in which the choices and problems in annotating these 25 texts were discussed

The purpose of the training sessions was to ensure uniform understanding of the principles underlying the abstract FTD categories, rather than to coach the annotators to get the answers right. The training sessions were followed by a proper annotation round using sets of 100-350 texts per annotator; each text received at least two independent annotations in each FTD. Table 3 presents the results of measuring the interval $\alpha$ values for the resulting annotations.

A final session was organised to discuss problems in annotating the texts, in particular, to detect cases when a text received no principal FTD or when two texts received exactly the same set of FTD values, while they were considered to be different in human judgement. This has led to new FTDs (**A16** to **A20**). For example, the appellative FTD (**A20**) has been introduced and labelled following (Jakobson, 1960) to reflect small ads, postings on dating websites, etc. Therefore, the full corpus presented in Table 1 has single annotations for these dimensions.

The average values of $\alpha$ in Table 3 are above .76. This indicates reasonable agreement, which is considerably higher than in many other interannotator agreement studies for genre annotation (Sharoff et al., 2010). Some FTDs, such as **A5** (entertaining) or **A6** (informal) in some studies demonstrated more disagreement. The annotation guidelines were deliberately based on text-external definitions by referring to the impression the annotators obtained from reading a text. The annotators noted that they could be more confident in choosing the value for **A6** if the guidelines had defined the linguistic features corresponding to informality, such as presence of slang. However, this omission was deliberate with the aim to help in differentiating text-external definitions from their text-internal linguistic realisations, so that at the next stage we can investigate the link between the FTD values and the linguistic features without circularity. Since the annotators differed in their perception of the degree of entertainment and informality, this led to lower $\alpha$ values in some cases.

The level of agreement also depends on the corpora used for annotation. The Pentaglossal corpus has been collected in a targeted way from parallel texts from a relatively small number of sources (243 texts from 14 sources). This usually increases the agreement values. Its Part 2 did not include any text which could be considered similar to

fiction, which led to total agreement on **A4** for this corpus. Another issue is that individual annotators could be biased towards certain interpretations of FTDs in different ways, thus lowering agreement in some annotation rounds. For example, in the GICR blogs study (Line 4 in Table 3) some annotators were biased in interpreting advertising texts (**A12**) or research texts (**A14**). Greater disagreement in a random selection of blogs can be also related to the fact that blogs are not subject to any gatekeeping. Therefore, some of them can be less conformant to the expectations an annotator has for a particular FTD. For example, a blog entry written by a researcher can express ideas in a form, which is different from what is acceptable in a research paper. Even if a blog entry discusses a research paper,[7] some annotators might prefer treating it just as an argumentative blog entry, which does not score on **A14**. However, the overall level of agreement for such categories is well above the acceptable threshold.

In one of the annotation rounds (GICR-random, Line 3 in Table 3) the annotators have been asked to annotate using the genre labels from (Egbert et al., 2015). This task was also embedded in the training session as yet another parameter for annotating texts. Even though there were far fewer problems in understanding the look'n'feel labels from Egbert&Biber in comparison to the more abstract FTDs, the annotators experienced problems in choosing the labels consistently. The nominal Krippendorff $\alpha$ was $0.53$ for agreement on the complete set of 56 labels and $0.66$ for agreement on the 8 higher-level categories ('narrative', 'instructional', 'description', etc). The most common examples of disagreement were 'Description of a thing' vs 'Encyclopedic article' and 'Personal diary blog' vs 'Opinion blog'. In the FTD framework such problems are avoided by having more general categories (**A16**, 'information') and the possibility to express the degree with which a text belongs to one category or the other (**A11** 'personal' vs **A17** 'reviews').

# 5 Analysis of corpora with respect to FTDs

In addition to a topological genre space which describes texts in 18 dimensions, it is interesting to investigate genres on a map, so that the distances between texts can be visualised in a two-dimensional space. One of the methods for dimensionality reduction is t-distributed Statistical Neighbour Embedding (t-SNE), which tries to map a higher dimensional space into a lower dimensional one by converting the distances between the data points in each space into similarity scores represented by conditional probabilities. The assumption is that the probability value is close to one for neighbouring data points and it is close to zero for widely separated ones. The lower dimensional space is adapted in iterations to minimise the Kullback-Leibler divergence between its probabilities and those from the higher dimensional space (Van der Maaten and Hinton, 2008).

It is difficult to compare different dimensionality reduction approaches using a metric, since each method optimises its own objective function, for example, orthogonalisation of linear combinations of the original dimensions in PCA or distortion of mapping in MDS. Empirically, t-SNE produced a two-dimensional representation explaining our data better in comparison to other approaches, including factor analysis (FA), Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Multi-

[7] For example `https://blog.kilgarriff.co.uk/?p=12`

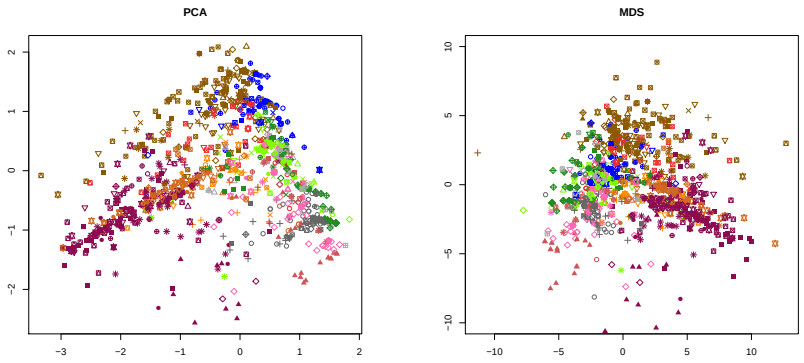


Figure 1: Comparison to PCA and MDS transforms

dimensional Scaling (MDS). Figure 1 shows examples output by PCA and MDS on our data. Linear methods, such as PCA, do not reflect non-linear interactions between the variables, while other non-linear methods, such as MDS, minimise the overall stress function for the distances between all objects. At the same time, t-SNE only preserves local distances between very similar objects. It does not penalise changes in distances between data points widely separated in the original dimension, since their similarity expressed via conditional probabilities is very close to zero anyway. In the end, the t-SNE map is easier to interpret, see Figure 2.

The colour labels in Figure 2 represent the **first** principal functional dimension of texts from the combined corpus reported in Table 1. The first FTD provides only one of several descriptors for a text in question, so some objects of the same colour can be further away in the map because of the values of other FTDs for these texts.

In order to determine how many traditional genres can be treated as common "syndrome" genres represented by a combination of FTDs, unsupervised clustering methods have been used, in particular, `pam` (partitioning around medoids), as implemented in the `cluster` package in R.

The quality of the obtained clustering solutions can be assessed with the silhouette method, i.e., ratio between the degree of similarity of objects within the same cluster vs those outside it (Kaufman and Rousseeuw, 2009). The best mean silhouette value has been obtained when the number of clusters is $17^{\pm 4}, sil \approx .55$, which indicates a fairly good separation of texts into clusters. We can assume that the useful number of labels to describe the genres of texts in terms of the chosen FTDs in this corpus is about 17. In Figure 2 the different clusters in the 17-clusters solution by `pam` are indicated by character shapes.

The solution with 17 clusters generated the following genres (with the number of their instances given in brackets):

**Cl1: A7** instructional texts, mostly FAQs (78);

**Cl2: A9** regulatory texts (65);

**Cl3: A17** reviews of products or services (53);

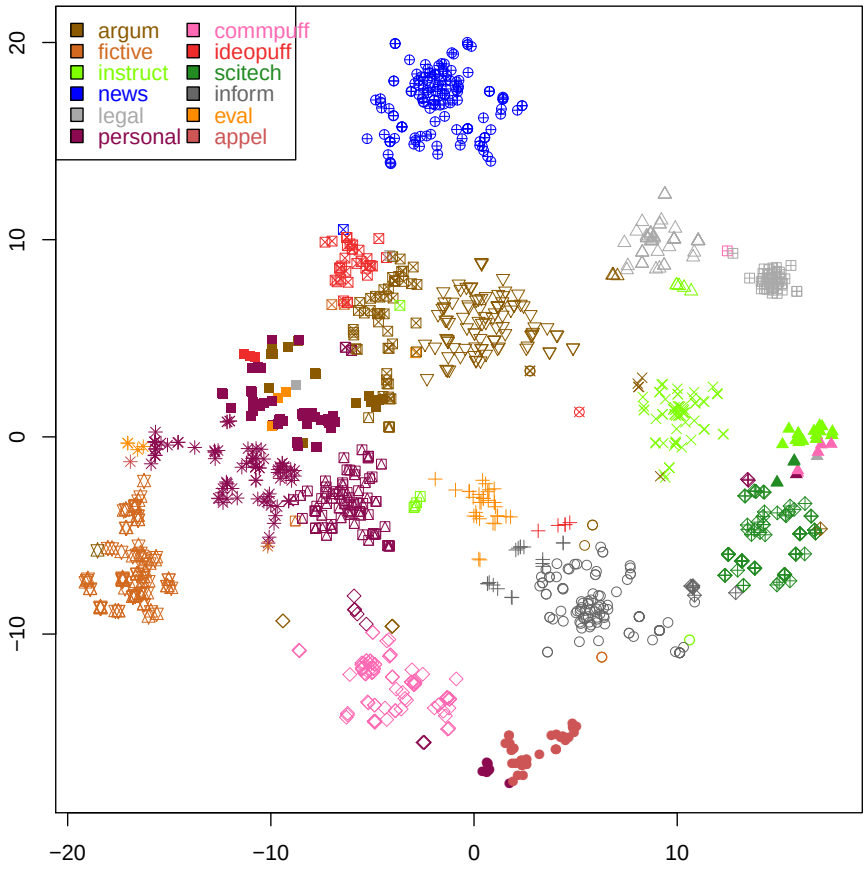


Figure 2: Locations of text-external FTD annotations via t-SNE transformation

**Cl4: A12** advertising (98);

**Cl5: A1** argumentative texts from newspapers and similar quality sources (124);

**Cl6: A16** texts aimed at presenting information, like Wikipedia articles (127);

**Cl7: A13 + A1** propaganda texts with some amount of argumentation, often coming from the websites of respective organisations (88);

**Cl8: A11** personal blogs, primarily objective reporting (108);

**Cl9: A14** research papers and similar research-related texts (92);

**Cl10: A8** hard news (124);

**Cl11: A4** fiction, some texts are also marked as **A5** 'entertaining' (69);

**Cl12: A9 + A16** legal summaries and clarifications (39);

**Cl13: A1 + A11** personal opinion blogs (43);

**Cl14: A11 + A3** personal diary blogs expressing one's emotions, some texts are marked as **A19**, attempts to embellish their style (38);

**Cl15: A11 + A17** personal blogs providing a review, the common topics are travel and shopping (117);

**Cl16: A7 + A14** academic instructions,[8] which are fairly distinct from other kinds of instructional texts, such as DIY manuals and FAQs (36);

**Cl17: A20** invitations to events or actions (50).

In the lists produced by `pam`, some of the clusters have a single dominant principal FTD. For example, hard news (**A8**) and fiction (**A4**) are the two clusters that can be described by a single FTD each, which indicates that they are presented as quite distinct in human perception. Other "syndrome" genres show a degree of variation in their FTDs. This is precisely the reason of their unreliable annotation in terms of atomic genre labels. The FTD corresponding to regulatory texts (**A9**) is detected reliably by the annotators. However, through clustering we can observe recurrent combinations with other FTDs. One of the clusters contains explanations and summaries for laws (**A16** inform, in addition to **A9**), another cluster of **A9** documents contains deliberations for court decisions (**A1**, argum), while another one presents instructions on how to act in a legally justifiable way, e.g., advice on paying taxes (**A7** instruct). Reporting one's personal experience (**A11**) in modern Web texts usually happens through the medium of blogs. In terms of the FTDs this is often combined with argumentation (**A1**), reviewing (**A17**), promotion (**A12**), instructions (**A7**), etc.

Clusters identified over the FTDs often correspond to traditional genres, such as those used in (Egbert et al., 2015). For example, 'News' as a genre label corresponds nicely to a cluster with **A8**. However, a text with a high **A8** value can be also argumentative (**A1**), light-hearted (**A5**), or it can contain an overview of a topic (**A16**). A traditional genre palette forces the annotators to choose one label, which can be 'News report', 'Short story', 'Magazine article', 'Opinion', 'Persuasive article', for their definitions see (Egbert et al., 2015). If the clusters or traditional genre labels are used, different annotators can naturally consider a text from different viewpoints, thus reducing reliability of their annotation. For example, the prototypical instances of CL10 contain primarily news reporting. However, in terms of Figure 2 CL10 texts which express opinions are located slightly lower than its core, closer to texts from either CL5 or CL13, depending on the amount of first-hand reporting. Further down the left side of Figure 2, diary-like blog entries (CL14) are located closer to fiction (CL11) and at some distance away from argumentative blogs (CL13), which are in their turn located closer to argumentative texts (CL5). The area of informative texts (CL6) in this map is situated between research papers (CL9) and reviews (CL3), thus reflecting natural variation of texts within this category.

[8] For example, `http://www.mml.cam.ac.uk/call/translation/toolkit/6/`

| | |
|---|---|
| Training docs | 503 |
| Features | 1035 |
| Correlation X | 0.86 |
| Correlation Y | 0.87 |

Table 4: Regression results with RVM

# 6 Learning the dimensions

The previous section offered genre analysis of a specific genre-annotated corpus. Ideally, we would like to determine the genre of any other text from the Web, in other words, to locate the position of any text on a map like Figure 2 or in the full FTD space. It is also important to compare different corpora with respect to their genre composition, for example, by showing the similarities and differences between the BNC and ukWac.

For a Machine Learning algorithm this means an attempt to use information obtained from text-internal linguistic features of texts in the training corpus to predict the coordinates of texts from another corpus. More specifically, we will use the known X and Y coordinates of subsets of our corpus for a language (English in this study) together with appropriate linguistic features to train a multivariate regression model. This study used features based on the frequencies of the 500 most frequent words (from the BNC) and the full set of POS tags. In addition, 23 features of the Biber tagger were included as implemented in (Wilson et al., 2010), for example, general emphatics or place adverbials.

Several multivariate regression methods are available in this situation. A commonly used one is SVM Regression with the RBF kernel (Smola and Schölkopf, 2004). However, its application to our corpus results in a large number of support vectors (about 400-450 for the 504 training instances, depending on the settings), which led to reasonable accuracy, but its application to Web data produced very inconsistent results since the model became overfit on the training data. Relevance Vector Machines (RVM) is a related formalism (Tipping, 2001), which builds support vectors from more representative (prototypical) samples in the training set, producing far fewer support vectors (60 to 120 in our corpus), which led to comparable accuracy and better stability on Web data. The correlation results obtained via 10-fold cross-validation are given in Table 4. This demonstrates that we predict the position of a text in the two-dimensional space with reasonable accuracy.

Once the ML models are available, they can be applied to a corpus with known categories, such as the Brown and BNC categories to map these corpora into the same two-dimensional space. In this way, we can compare the similarities and differences between such corpora using known genre categories and samples, even if the sets of genre categories in the two corpora are not compatible.

Figure 3 presents the density map for the texts in the Brown Corpus (BC) together with the median positions of its respective categories as well as the median positions of the 25 most frequent genre categories from the BNC, excluding `W_misc` (the BNC categories start with `W_` for written texts and `S_` for spoken sources). The density map is another contribution of the topological perspective on genre classification, as it shows the position of the majority of texts in a corpus. This can be compared against the

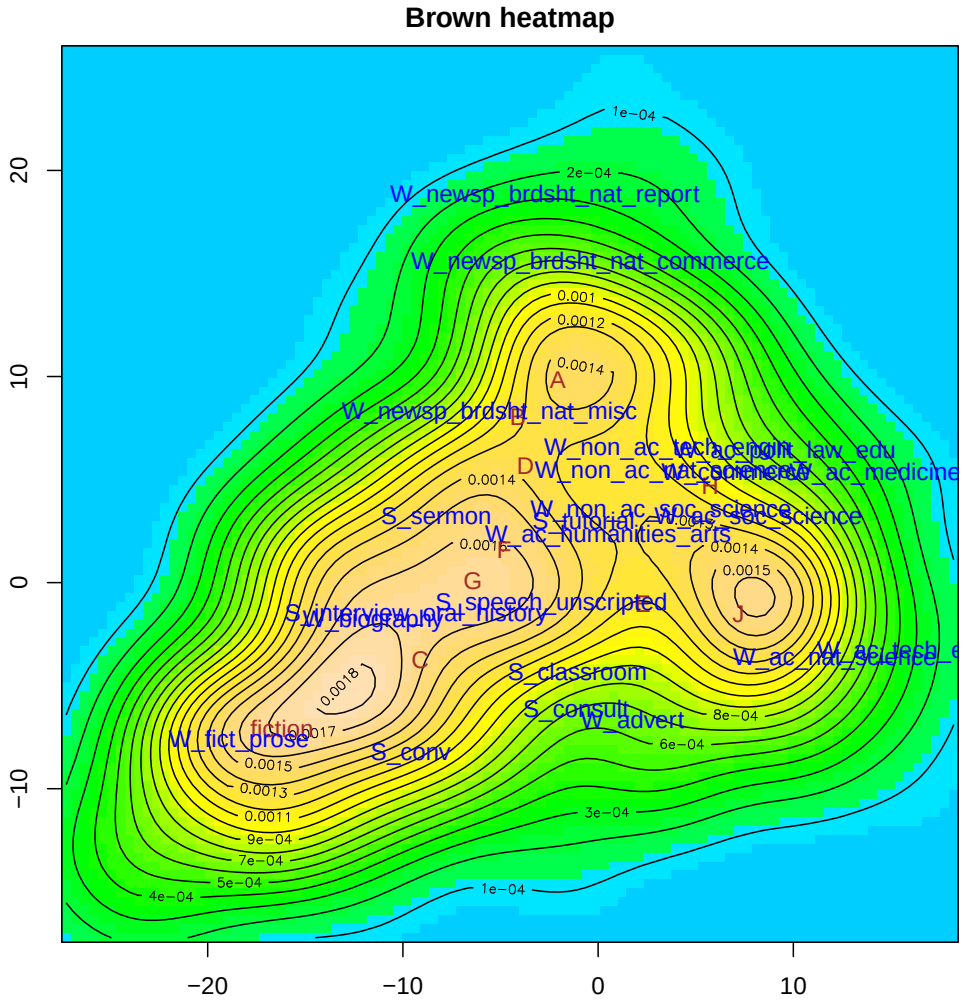


Figure 3: Locations of genres in Brown Corpus and BNC

position of the majority of texts in each category. The map in Figure 3 confirms that the three most commonly used text types in BC correspond to newspapers, fiction and research articles.

The fiction cluster demonstrates a clear identity across both corpora, and it neatly coincides with the position occupied by fiction in Figure 2. Other genre positions vary depending on variation within their constituent text. The Category J ('learned') texts in BC together with several `W_ac_` categories from the BNC also occupy similar positions, and they match the research and information clusters from Figure 2. However, the positions of individual texts within this category vary considerably. For example, Text J31 from the Brown Corpus is located on the opposite side of the map ($x = -11.7, y = -11.2$), in the same position as personal diary blog reports in Figure 2, which can be explained by the fact that this research article predominantly consists of reporting individual cases:

| # | N:49,782 | # | S:102,248 | # | F:2,508 | # | A:211,301 |
|---|---|---|---|---|---|---|---|
| 685 | guardian.co.uk | 1232 | cam.ac.uk | 417 | classic-literature.co.uk | 579 | travelpublishing.co.uk |
| 506 | demon.co.uk | 1112 | ox.ac.uk | 121 | ex.ac.uk | 402 | bookweaver.co.uk |
| 290 | acronym.org.uk | 979 | ed.ac.uk | 75 | athelstane.co.uk | 386 | pwp.blueyonder.co.uk |
| 234 | iwar.org.uk | 554 | leeds.ac.uk | 41 | demon.co.uk | 372 | www.faber.co.uk |
| 216 | independent.co.uk | 554 | demon.co.uk | 31 | omnia.co.uk | 361 | www.fpigraphics.co.uk |
| 215 | cam.ac.uk | 445 | bham.ac.uk | 26 | tvradiobits.co.uk | 289 | classic-literature.co.uk |
| 200 | poptel.org.uk | 444 | ucl.ac.uk | 21 | pandemonium.me.uk | 278 | chrysalisbooks.co.uk |
| 191 | greenparty.org.uk | 406 | open.ac.uk | 20 | guardian.co.uk | 241 | directline-holidays.co.uk |
| 185 | icnetwork.co.uk | 403 | gla.ac.uk | 19 | weddingguide.co.uk | 236 | londonpropertywatch.co.uk |
| 172 | webstar.co.uk | 340 | soton.ac.uk | 19 | applesofgold.co.uk | 233 | www.chycor.co.uk |
| 145 | indymedia.org.uk | 307 | man.ac.uk | 18 | myarseisnotpansy.co.uk | 228 | www.vam.ac.uk |

Table 5: Areas of ukWac

> One hebephrenic woman often became submerged in what felt to me like a somehow phony experience of pseudo-emotion, during which, despite her wracking sobs and streaming cheeks, I felt only a cold annoyance with her. Eventually such incidents became more sporadic, and more sharply demarcated from her day-after-day behavior, and in one particular session, after several minutes of such behavior - which, as usual, went on without any accompanying words from her - she asked, eagerly, "Did you see Granny"? At first I did not know what she meant; I thought she must be ...

Similar variation within the category applies to the texts in the reportage cluster (Category A). Overall, the Category A texts in BC are positioned lower in the vertical dimension in comparison to the news texts in the training set and also lower than purely reporting texts from the BNC. For example, Text A04 discussed above is positioned at $x = -1.3, y = 13.1$, which is closer to argumentative texts than to hard news.

This compares the genres of the BC and BNC texts. However, the goal of this study is to provide a way of assessing the composition of Internet corpora, so that these corpora can be compared to existing genre inventories. Another goal is to provide the possibility of selecting subsets of Internet corpora. In the absence of gold-standard data, one way of interpreting the results of classification on the entire ukWac is by selecting areas of interests and by extracting the most common domain names from them.

Following the locations of documents in Figures 2 and 3, four areas of interest have been selected as $\pm 3$ around the following points:

**N** $x = 0, y = 15$ This should be similar to hard news;

**S** $x = 10, y = 0$ This should correspond to traditional research papers, mostly in hard sciences;

**F** $x = -20, y = -10$ This should correspond to fiction and similar narrative texts;

**A** $x = 0, y = -15$ This should correspond to advertising;

Table 5 summarises the results for each area of the ukWac in terms of their most common URLs, the number of documents in each area is given after its name. As

can be expected, the **N** area primarily contains pages from news sources. Apart from *The Guardian* and *The Independent*, other major British newspapers have restrictive crawling strategies, so they have not been collected into ukWac. Therefore, this area is populated by news items from such sources as the Campaign Against Sanctions on Iraq (`casi.org.uk`), the Acronym institute, or university newsletters. The science domain is quite predictably dominated by descriptions of research projects from the major British universities. While fiction is under-represented on the Web in comparison to other genres, the fiction area is populated by out-of-copyright fiction (e.g. `classic-literature.co.uk`), as well as fan fiction from web-hosting providers (`demon.co.uk` or `omnia.co.uk`). The texts of the advertising region cover the topics of renting a house, travel, gaming, as well as book advertising, for example:[9]

> The fast, easy way to master the essentials of Spanish. Now, learning Spanish can be as easy as uno, dos, tres! Combining the quick-reference virtues of a phrase book with the learning tools of a full-fledged language course, this popular guide gives you a solid start ...

# 7 Conclusions and further work

First, this study proposes a text-external framework of Functional Text Dimensions for assessing human perception of the similarities between texts in terms of their function. The annotation scheme is intended to 'say sensible and useful things' about the majority of texts from general-purpose corpora derived from the Web. In our annotated set of 1,400 texts, similar texts received similar annotations, while texts judged to be considerably different received annotations which differed in at least one FTD.

Double-annotation experiments also show that agreement on the FTDs is higher than what is usually achieved with atomic genre labels. The higher agreement leads to better accuracy of automatic classifiers and to more interpretable results of comparison between the categories. To visualise the multidimensional text-external space as a genre map, this space has been presented as a two-dimensional representation using t-distributed Stochastic Neighbour Embedding. A qualitative investigation of this space shows the more typical clusters corresponding to genres as "syndromes" of the FTDs. The corpus is derived from Web texts, so the common genres detected in this way are predominantly specific to the Web. Still many of them are directly comparable to what can be found in traditional corpora.

Second, a text-internal Machine Learning framework has been developed, which can map any text into the text-external representation produced in the first step. In the end, different corpora can be compared against a common set of categories, for example, text types of the BNC can be compared to those in the Brown corpus. Some text types map to each other across corpora consistently, for example, fiction. The position of some other text types on the map can differ because of occasional differences in the composition of the respective corpora. For example, this concerns news, as Category A in the Brown Corpus is more argumentative than the respective subcategories of `W_newsp` in the BNC. The framework can also account for individual differences

[9] `http://encyclopedia.classic-literature.co.uk/dictionary-store/0471134465`

between the texts within the same genre, such as the 'learned' texts (Category J) in the Brown Corpus. Similarly, a Web-derived corpus can be investigated in the same map by considering its reasonably sized subcorpora corresponding to the same regions as in the traditional corpora.

More generally, this defines a framework for detecting the similarity in terms of genres between texts via their position in the space of either two or eighteen dimensions. The results of annotation of the texts reported in this paper, as well as the tools for processing any further corpora are available under permissive licenses.[10] This should help in replication of the results of this study, as well as in annotating new corpora in the FTD framework.

One of the directions for further research concerns development of a bigger and more varied corpus. Even though, the corpus used for these experiments contain about 1,400 texts, more than one third of the BNC, the number of texts in each individual language is considerably smaller, about 1150 for Russian, 500 for English, 246 for Chinese and French (because 246 texts from the Pentaglossal corpus are available in their translations). It is also promising to study text types not only across corpora, but also across languages, so that other large corpora, e.g., those from the WAC family (Baroni et al., 2009), can be compared against each other over the same map.

Another direction concerns selection of the text-internal features for text classification. The basic set used in this study achieves reasonable results with correlation in regression tasks for the two-dimensional map above 0.85. However, the correlation results for the individual FTDs is much less satisfactory, down to 0.20 for some of them, for example, for informative (**A16**) and appellative (**A20**) texts. This suggests that better Machine Learning methods are needed.

One of the limitations of the study is that the classifier for the experiments reported in Section 6 was trained on an annotated web corpus and was applied to data from traditional corpora. This led to interpretable results, because texts which are similar in their function across these corpora are positioned in the same areas, also comparable to the annotated Web corpus. However, ideally we need to adapt the classifier to the genres and linguistic features of traditional corpora in a more flexible way, for example by using Self-Taught Learning (Raina et al., 2007).

Another direction of research concerns studying a link between text-external categories and their text-internal linguistic realisations. One possibility to achieve this is by using Biber's Multi-Dimensional Analysis, so that the text-external categories ('registers' in Biber's terminology) are provided by the FTDs, while a set of linguistic features, such as those used in (Biber, 1995), can be investigated via Factor Analysis to reveal their link to the text-external categories.

## Acknowledgements

I'm grateful to Richard Forsyth who initially suggested the idea for decomposing the genre labels used in FGC into testable symptoms. I'm also grateful to Anisa Katinskaya, Adam Kilgarriff, Vladimir Selegey, Alexey Sorokin for extensive discussions

[10] `http://corpus.leeds.ac.uk/serge/webgenres/`

and help in organising the annotation experiment. Special thanks to the reviewers for their suggestions. The usual disclaimers apply.

## References

Adamzik, K. (1995). *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

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

Baroni, M., Bernardini, S., Ferraresi, A., and 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.

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Crowston, K., Kwasnik, B., and Rubleske, J. (2010). Problems in the use-centered development of a taxonomy of web genres. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer.

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

Forsyth, R. and Sharoff, S. (2014). Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, 29:6–22.

Görlach, M. (2004). *Text types and the history of English*. Walter de Gruyter.

Grieve, J., Biber, D., Friginal, E., and Nekrasova, T. (2010). Variation among blogs: A multi-dimensional analysis. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Halliday, M. A. (1985). Register variation. In Halliday, M. A. and Hasan, R., editors, *Language, context, and text: Aspects of language in a social-semiotic perspective*, pages 29–41. Oxford University Press.

Jakobson, R. (1960). Linguistics and poetics. In Sebeok, T. A., editor, *Style in Language*, pages 350–377. MIT Press.

Katinskaya, A. and Sharoff, S. (2015). Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proc BSNLP'15*, Sofia.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3).

Kulas, J. T., Stachowski, A. A., and Haynes, B. A. (2008). Middle response functioning in likert-responses to personality items. *Journal of Business and Psychology*, 22(3):251–259.

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

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Lemke, J. L. (1999). Typology, topology, topography: genre semantics. MS, University of Michigan.

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., and Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In *Proc* $8^{th}$ *Web as Corpus Workshop (WAC-8)*.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: Transfer learning from unlabeled data. In *Proc.* $24^{th}$ *International Conference on Machine Learning*, ICML '07, pages 759–766, New York, NY, USA.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sharoff, S. (2005). Methods and tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., and Rayson, P., editors, *Corpus Linguistics Around the World*, pages 167–180. Rodopi, Amsterdam.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 149–166. Springer, Berlin/New York.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC 2010*, Malta.

Sinclair, J. and Ball, J. (1996). Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Sorokin, A., Katinskaya, A., and Sharoff, S. (2014). Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo.

Tipping, M. E. (2001). Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Waller, R. (2012). Graphic literacies for a digital age: the survival of layout. *The Information Society*, 28(4):236–252.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proc 47th Annual Meeting of the ACL*, pages 674–682.

Wilson, J., Hartley, A., Sharoff, S., and Stephenson, P. (2010). Advanced corpus solutions for humanities researchers. In *Proc Advanced Corpus Solutions, PACLIC 24*, pages 36–43, Tohoku University.

# Appendix 1: Functional Text Dimensions

| Code | Label | Question to be answered |
|---|---|---|
| **A1** | argum | To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', for argumentative blogs, editorials or opinion pieces) |
| A3 | emotive | To what extent is the text concerned with expressing feelings or emotions? ('None' for neutral explanations, descriptions and/or reportage.) |
| **A4** | fictive | To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.) |
| A5 | flippant | To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader? ('None' if it appears earnest or serious; even when it tries to keep the reader interested and involved) |
| A6 | informal | To what extent is the text's content written in a fairly informal manner? (as opposed to the "standard" or "prestige" variety of language) |
| **A7** | instruct | To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ) |
| **A8** | hardnews | To what extent does the text appear to be an informative report of events recent at the time of writing? (Information about future events can be hardnews too. 'None' if a news article only discusses a state of affairs). |
| **A9** | legal | To what extent does the text lay down a contract or specify a set of regulations? (This includes copyright notices.) |
| **A11** | personal | To what extent does the text report from a first-person point of view? (For example, a diary-like blog entry.) |
| **A12** | compuff | To what extent does the text promote a product or service? |
| **A13** | ideopuff | To what extent is the text intended to promote a political movement, party, religious faith or other non-commercial cause? |
| **A14** | scitech | To what extent would you consider the text as representing research? (It does not have to be a research paper. For example, 'Strongly' or 'Partly' if a newswire text has scientific contents.) |
| A15 | specialist | To what extent does the text require background knowledge or access to a reference source of a specialised subject area in order to be comprehensible? (such as wouldn't be expected of the so-called "general reader") |
| **A16** | info | To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books). |
| **A17** | eval | To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, by providing a product review). |
| A18 | dialogue | To what extent does the text contain active interaction between several participants? (For example, forums or scripted dialogues). |
| A19 | poetic | To what extent does the author of the text pay attention to its aesthetic appearance? ('Strongly' for poetry, language experiments, uses of language for art purposes). |
| **A20** | appell | To what extent does the text requests an action from the reader? ('Strongly' for requests and other appellative texts). |

The following levels are used:

| **Rating** | **Levels:** |
|---|---|
| 0 | none or hardly at all; |
| 0.5 | slightly; |
| 1 | somewhat or partly; |
| 2 | strongly or very much so. |

A text is expected to have a **Strongly** annotation for at least one dimension given in bold. For some texts it is possible to have several **Strongly** annotations, but each choice needs to be justified by how strong the dimension is present in this text.

The numbering of the FTDs listed above is not consecutive because of the need to keep compatibility with previous studies. Specifically, **A2** and **A10** have been deleted to simplify annotation.