

Guidelines on genre annotation of webpages

Serge Sharoff

Centre for Translation Studies, University of Leeds

1 Taxonomy

For assessing the composition and for comparing large corpora we need a genre taxonomy that is 1) broad, i.e., it can be applied to any text, and 2) compact, i.e., it contains a small number of labels, so that any assessment or comparison is done according to an observable number of parameters. More information on the rationale behind the proposed set of labels is available in (Sharoff, 2009).

Your task is to annotate a set of webpages according to the following labels:

1. **information** – catalogues, glossaries, as well as purely informative texts like CVs, homepages, specifications or encyclopedic factsheets;
2. **instruction** – how-tos, FAQs, tutorials;
3. **propaganda** – adverts, political pamphlets;
4. **recreation** – fiction and popular lore (this also includes narrative biographies and memoirs);
5. **regulations** – laws, small print, contracts;
6. **reporting** – factual texts reporting on a state of affairs, like newswires (including sport) and police reports;
7. **discussion** – all texts expressing positions and discussing a state of affairs, the three main subtypes are **public** (corresponding to public debates, like blogs or opinionated journalistic texts), **academic** (research papers, books), and **communication** (spontaneous electronic communication, like discussion forums or chat rooms);
8. **unknown** – this was reserved for webpages with little or no running text, like forms for queries, logins, download pages, flash animation, samples of source code, etc; one important subcategory here is **index**, i.e., portals, sitemaps, other lists of links (mostly containing incomplete or isolated sentences).

Each label in this set corresponds to a generalised aim of text production, i.e., **instruction** is for texts aimed at teaching how to achieve something, **recreation** is written for leisure-time reading. If a text cannot be comfortably classified as 1-6, it can be safely considered as **discussion**, unless it does not contain running text. If more than one label can be applied, feel free to do this.

2 Explanations for specific labels

2.1 Information

This label is aimed at texts *only* providing information, such as catalogues, lists of people, places, businesses, objects, dictionary definitions or encyclopedic articles on a subject, short summaries of longer texts, minutes of meetings, etc. If a page can be classified using other categories, use them instead of **information**.

2.2 Instruction

These texts are aimed at teaching or explaining how to achieve something. The majority of texts classified with this label belong to two types:

- structured lists, such as FAQs, recipes, lesson plans, steps for assembling, repairing or maintaining something (often such texts contain imperatives);
- advice written in a more narrative style, such as a recommendations on how to choose something, tutorials.

Some (!) research articles can fall under this label if their main purpose is to give direct advice, not to analyse an issue.¹

2.3 Propaganda

These texts are aimed at selling a product, service or political opinion. Often such texts praise the advantages of using the service or supporting a person. However, if an opinionated political text also discusses an issue (not just saying *X is the best*), it is better to classify it as **discussion**.

2.4 Recreation

There is a large variety of texts aimed at recreational reading, such as fiction, popular lore, narrative biographies and memoirs. However, the category is not designed to include **discussions** pertaining to entertainment, e.g., interviews with celebrities, travelogues, film reviews, etc.

2.5 Regulations

Texts classified in this way correspond to various rules, bills, small-print, contracts or official agreements.²

¹e.g., http://www.privcom.gc.ca/media/nr-c/opinion_021122_lf_e.asp

²e.g., <http://contracts.onecle.com/talk/walsh.nso.2000.08.07.shtml>

2.6 Reporting

This label applies to factual texts describing what has happened. In addition to newswires and police reports, the category includes factual descriptions of historic events. However, this label is *not* designed for more narrative entertaining biographies, like those of pop stars or spy stories; they are better classified as **recreation**. If there is a discussion of possible reasons behind the event, it is safer to classify it as **discussion**. **discussion** also applies to diary-like blog entries.

2.7 Discussion

This is the biggest category with a variety of subtypes. The three main subtypes are **public** (corresponding to public debates, like blogs or opinionated journalistic texts), **academic** (research papers, books), and **communication** (spontaneous electronic communication, like discussion forums or chat rooms). Take care when selecting the subtypes. A diary-like blog entry is normally aimed at **communication**, while an argumentative political blog is considered to be a **public** discussion.

2.8 Unknown

This label is for webpages which are *not* designed to be read like a text. This includes three main categories, **interaction**, e.g., forms for queries, logins, download pages, **nontext**, e.g., flash animation, videos, samples of source code, and **index**, e.g., portals, sitemaps, other lists of links (mostly containing incomplete or isolated sentences).

3 Dealing with the ambiguity

Genres are often defined as a triad of `<form, content, function>`. Since content is part of the definition of genres, sometimes it is difficult to ignore contributions coming from the content. Nevertheless, in this coding exercise it is important to annotate according to the main aim of text production. To take the legal domain as an example: not all legal texts are **regulations**. A law has to be coded as **regulations**, but a report about a legal case is **reporting**, advice on how to act in a lawsuit is **instruction**, a research article on a legal issue is **discussion**.

There is some confusion with coding narratives, e.g., biographies or historical accounts. The rule of thumb is based on the main aim of the text in question: if it looks like a factual account of what happened, it is **reporting**, if it looks like a text written for entertainment, it is **recreation**. However, the genre of social gossips (*Prince Andrew wants to divorce*) is not different from news (*President Bush wants to pass a new bill*), so it is still considered as **reporting**. For sports news and gossips, you should also use **reporting**. However, a story in

“Cosmopolitan” about the life of Prince Andrew (or President Bush) mentioning his thoughts or private events will be **recreation**.

The degree of **propaganda** in a text is often in the eyes of the annotator. Often a home page tries to show the individual or the company it belongs to in the best possible way. Normally, such texts are still classified as **information**, unless they actively promote services offered by this individual or company.

In some cases, different parts of a text belong to different categories, e.g., a text might be an **instruction** on how to do X, while also giving **information** on how X relates to Y. In such cases it is better to annotate according to the aim of the biggest part and leave a comment in the third column.

References

Sharoff, S. (2009). In the garden and in the jungle. Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., Rehm, G., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.