COWCat document categorization guidelines 2016

Roland Schäfer
Linguistic Web Characterization (DFG SCHA1916/1-1)
Freie Universität Berlin

March 6, 2016

The following people contributed to the classification scheme and the CatTLE reference corpus:
Felix Bildhauer, Sarah Dietzfelbinger, Lea Helmers, Theresia Lehner, Kim Maser, Samuel Reichert, Luise Rißmann

1 Background

These guidelines are based to a large extent on Sharoff (2006), which itself is based on an EAGLES specification: `http://wackybook.sslmit.unibo.it/` The guidelines were adapted for the COW project: `http://corporafromtheweb.org`

2 General Instructions

- Only the text that is in the actual corpus documents should be considered. If the full HTML page contains more material, which was accidentally removed by the post-processing software, this material should be ignored.
- Documents which are spam, non-coherent, nonsense, or contain significant amounts of HTML or script code due to post-processing errors should be marked as **faulty**.
- Each document which is not faulty must be coded by selecting values in each category. There must not be any blank or N/A fields.
- Note: *Audience* was dropped in 2016 because of consistently incoherent annotation results.

3 The taxonomy

3.1 Authorship (At)

To determine the value for Authorship (especially *Single, female* and *Single, male*), external material such as the original web page can and should be taken into consideration. Inferring authorship (including gender) from obvious personal names within host names (such as *felixbildhauer.blogspot.com*) is admissible.

1. **Single, female (Sf)**
   The document was written by a single female author. *Female* refers to gender, not sex.
2. **Single, male (Sm)**
   The document was written by a single male author. *Male* refers to gender, not sex.

3. **Multiple (Mu)**
   The document was written by multiple authors, either collectively or in some form of dialogue. Blog posts with comments (even a single comment) count as *Multiple*. A forum thread with only one poster is not *Multiple*.

4. **Corporate (Co)**
   Any text written in the name and interest of a company, association, party, syndicate, etc. A typical lexical indicator is the use of first person plural pronouns. A text signed by a single high-ranked representative (president, CEO, etc.) on the web page of a company is *Corporate* unless the text is about private matters (personal illness, own family, etc.). All statutes, bylaws, regulations, and laws are always *Corporate*. Also, a project specification for a scientific project (grant application or description) is usually best coded as *Corporate*.

5. **Unknown (Un)**
   None of the above.

3.2 Mode (Mo)

Annotators can assign one, two, or all three modes to mixed documents.

1. **Written (Wr)**
   A document which which was written as a traditional text (or a collection of such texts), with some recognizable plan and structuring, most likely having undergone at least some editing. There is no minimal length for *Written*, even a Haiku is *Written*.

2. **Spoken (Sp)**
   Any text that is clearly marked as spoken. Usually, in web corpora, this is rare. Typical cases are transcripts of speeches or interviews, even if they look like they have been heavily edited in the transcription process.

3. **Quasi-Spontaneous (Qs)**
   Text that was written in a spontaneous act. Such texts look unedited and include spelling errors, missing punctuation, non-sentential utterances like exclamations. They often contain non-standard grammar (closer to spoken language) and orthography as well as slang vocabulary. Typographically, smileys and other emoticons are typical markers of *Quasi-Spontaneous*. Such documents usually involve some form of dialogue. Typically, they are chat transcripts/logs and unmoderated spontaneous discussions in forums.

3.3 Aim (Ai)

Annotators can assign one or two aims to mixed documents.

1. **Recommendation (Re)**
   Any text which advertises a certain product or service in the interest of the author(s). The recommendation may be motivated by, for example: commercial, religious, or political interest/belief or legal requirements. Not *Recommendation*, however, is any review of a product (like a book or a movie) written by an independent person; this is *Discussion*. Also, vague recommendations as to how one should live her/his life, etc. are rather *Instruction* or even *Discussion* (cf. directly below).

2. **Instruction (Is)**
   Any instructions on how to do or achieve something, uttered in a helping spirit. For example: An FAQ, health recommendation, teaching materials, spiritual instructions on how to achieve enlightenment, etc. A forum thread where someone asks for advice and possible solutions are discussed is *Discussion*, not *Instruction*. Also, if some (e. g., legal) situation is explained, and instructions can

– in principle – be inferred indirectly from the explanation, annotators should code *Information*. Finally, if the advice is not coming from the author, but the author rather reports about someone giving advice, the document should be coded as *Information* instead. Linguistically, imperatives or similar constructions (*do this, do that,...*) are good indicators of *Instruction*. Statements and conditionals are indicators of *Information*.

3. **Information (If)**
   Any text that provides statements of facts without any personal involvement or statement by the author(s). For example: science reports, statistical reports, news items which contain no commentary at all (which should be rare for scientific papers as opposed to simple scientific data reports). However, as soon as a report contains valuations like *a beautiful leisure resort*, it is rather *Information* but *Discussion*. All teaching materials are *Information*. A grant application is *Information* by definition.

4. **Discussion (Di)**
   Any text where authors report their own views or interpretations of facts or engage in a discussion thereof. For example: political or social comment, book or product reviews, blog and forum discussions, sermons, etc. Most scientific/scholarly papers contain some kind of comment or discussion and are most likely *Discussion*.

5. **Fiction (Fi)**
   All fictional text, including poems.

3.4 Domain (Do)

A topic is *what a document is about* in a narrow sense. A topic can be inferred *internally* from the characteristic vocabulary used in the document. A topic domain, on the other hand, is a cluster of topics that need not necessarily have a huge overlap w. r. t. their characteristic vocabulary. The definitions of topic domains rely to some extend on *external* knowledge about topics that belong together. For example, a document about Angela Merkel's view on the refugee crisis (the document's topic) and a document about Barak Obama's view on the financial crisis might not share a large vocabulary, but they are both about topics from the domain of *politics*.

We expect that a substantial proportion of the documents belongs to more than one domain simultaneously. Therefore, annotators should assign between one and four domains to each document. The domains assigned to a document can be weighted. The weighting is implemented by distributing a total of 4 points between the domains. For example, a document can receive 2 points (50%) for *Politics*, 1 point (25%) for *Society*, and 1 point (25%) for *Infrastructure*. There are no restrictions on the distributions, and anything between 4/0/0/0 (document belongs to one domain exclusively) and 1/1/1/1 (document belongs to four domains with equal weights) can be assigned.

1. **Science (Sc)**
   Anything related to empirical fundamental science. Primarily, Mathematics as well as Engineering and Natural and Behavioral Sciences, including linguistics. The document does not have to be written for an academic audience. Notice that reports about education (schools and universities as institutions, job education) are not automatically *Science* if they are not about actual research. If they are more about organizational, political and institutional matters, they might (additionally or exclusively) be, for example, *Politics* etc.

   Science does not include

   - Medical Science (cf. *Medicine*)
   - Political Science, Sociology, Macroeconomics (cf. *Politics* and *Society*)
   - Historical Science (cf. *History*)

- Microeconomics (cf. *Business*)
- Jurisprudence (cf. *Law*)
- Literary Science, Science of Art, Musicology (cf. *Arts*)
- Philosophy (cf. *Philosophy*)
- Theology (cf. *Beliefs*)

2. **Technology (Te)**
   Anything from information science to programming language tutorials, reports about PC hardware, internet technology, etc. Also biotech, military technology, space and aircraft technology, reports about technology/machinery relevant to certain professions (like agriculture, building or plumbing) etc. Even a document on diverse types of wood as a construction material is *Technology*. If the text is about fundamental science related to some technology (i. e., not applied), cross-classify with *Science*.

3. **Medicine (Me)**
   Medical advice, discussion of personal medical problems, nutrition advice, diets, medical science, genetics. This is not for discussion of methods not accepted by medical science such as reiki, faith healing, homeopathy, etc. These are *Beliefs*.

4. **History (Hi)**
   Anything related to history. This is in many cases *Politics* for the past. History ends roughly at 1990, such that the breakdown of the USSR, German reunification, the Gulf Wars, the Yugoslav Wars, the end of Apartheid in South Africa, etc. are not yet history.

5. **Philosophy (Ph)**
   Anything related to Philosophy. Philosophy is not empirical (like *Science*), and it is not bound by a specific religious view of the world (like *Beliefs*).

6. **Beliefs (Be)**
   Anything related to religion, the after-life, the supernatural, UFOs, etc. This includes anything from theological dissertations to speculative forum posts about reincarnation and previous lives, Area 51 conspiracies, etc. If the subject is a healing method not accepted by the general medical profession (including borderline cases like hypnosis), code *Beliefs*. If the document also includes an evaluation by medical experts, cross-classify with *Medical*.

7. **Law (Lw)**
   Anything related primarily to legal matters. This includes discussion of personal legal affairs, but also trials at the European Court of Human Rights. A report about the constitution of a country which is not yet in effect (i. e., still under discussion) is *Politics*, not *Law*. Once it has been officially ratified and there is discussion about its interpretations, a report about the interpretation is *Politics*. Statutes, bylaws, regulations, and laws are always *Law*.

8. **Politics (Po)**
   Anything related to current politics. This includes any political discussion about education, universities, the job market, international relations, war, etc. Mere administrative information on/ discussion about schools and universities, job agencies, etc. are *Infrastructure*, however. Reports about the second Gulf War, unemployment in the U.S., the Euro Crisis, the revolution in Egypt, financial problems of the Swedish health care system, the University of Göttingen losing its status as *Exzellenzuniversität*, etc. are *Politics*. Environmental issues under a political perspective are also *Politics*.

9. **Society (So)**
   Anything related to whole societies (or at least large parts of societies), but which is not related to

politics. In other words, anything that has an impact on the society but is not paid by or related to taxes. For example, environmental organizations and other NGOs, labor unions, non-political demonstrations, health insurance (if not tax-paid). *Society* can, of course, be cross-classified with *Politics*.

10. **Infrastructure (Ir)**
    Anything related to official authorities (fire brigade, police, military, public departments, schools, etc.) or state organized infrastructure. Discussions about road construction, practical advantages of fire drills at schools, train delays, job agencies. This also includes administrative information by schools and universities (organization of study programs, how to find an internship, etc.). If there is additional political or social impact, select an appropriate additional domain.

11. **Business (Bi)**
    Anything related primarily to business in the sense of microeconomics (not macroeconomics), primarily companies of any size. This includes discussion of personal economic affairs. For example also discussion of economic problems connected with personal plans of emigration to another country or the annual report of a company.

12. **Public (Pu)**
    Anything related to associations or clubs that have no political or socio-political goals and no larger social impact and no commercial or religious goals. For example, associations of stamp collectors, privately organized day-care organizations, study groups. A typical document reports on the activities of such associations as a whole (such as an annual meeting).

13. **Incident (Ic) [formerly: Life]**
    An incident that occurs in everyday life and intrinsically only affects a restricted number of people. For example local crime and accidents.

14. **Nature (Na)**
    Descriptive reports about natural phenomena (cross-classify with *Science* if a report is both descriptive and scientific). Wildlife, flowers and plants, climate, weather, earthquakes, meteorites, phenomena on other heavenly bodies, etc. Agriculture is not *Nature* unless the report is about effects of agriculture on nature or v. v.

15. **Sports (Sp)**
    Anything about professional sports or sports events worthy of media coverage. Pro soccer, football, darts, snooker, weight lifting, etc.

16. **Arts (Ar)**
    Discussion about arts and music; from painting, music, theatre to street art, hip hop battles, etc. The quality of the discussed works of art is irrelevant. The works of Bob Ross and Justin Bieber are as much *Arts* as those of Ernst-Ludwig Kirchner and Paul Hindemith. Includes literary science but not fiction itself (which is *Fiction*). Games (including computer games) are not *Arts* but *Leisure*.

17. **Entertainment (En)**
    Reports and discussion pertaining to the entertainment industry but not the actual output of that industry (which is *Arts*). Drug problems, love affairs, fashion choices, financial problems etc. of pop artists, actors, famous piano players, etc. This includes talk about royalty in case they have no ruling/governing authority in the respective country (England, Netherlands, Scandinavia, Spain, etc.).

18. **Leisure (Le)**
    Anything about personal hobbies or other leisurely activities. Personal sports and work outs, going out (to the theatre, a restaurant), pets, fashion, gaming, gambling, travelling, porn, gardening, cookery, etc.

19. **Individual (Iv)**
    Any document centered around a single person, like a biography or autobiography, an obituary, a laudation, etc. Self-portraits and CVs or documents containing someone's personal views on a collection of diverse matters without clear focus are *Individual*.

20. **Fiction (Fi)**
    Reports of fictional events, literary or not. Cross-classify if possible. For example, narrative passages from the Bible are typically at least *Fiction* and *Beliefs*. As yet another example, the first part of the Tin Drum could be *History* and *Fiction* because it is about fictional events but with a backdrop of true historical events.