COW text categorization guidelines

The FU Berlin COW Staff

March 2, 2012

These guidelines are a work in progress. Please always refer to the current version as found on the project home page.

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://hpsg.fu-berlin.de/cow/`

2 The taxonomy

2.1 Hierarchically interpreted categories

Some categories should be interpreted hierarchically in case of problematic decisions. I. e., if a coder cannot decide for *Mo*, *Au*, *Ai*, or *Do* which value to code, s/he should code the value (of the multiple applicable ones) which is highest in the category list. For example, a discussion of a new Google product including technical aspects *Do:Te* as well as legal and economic aspects *Do:Bl* is *Do:Te*, because *Do:Te* is higher in the list of values for *Do*. Of course, no valuation is implied with the hierarchy.

2.2 Authorship (At)

Note: Inferring authorship (including gender) from obvious personal names within host names (such as *felix-bildhauer.blogspot.com*) is admissible.

1. **Single, female (Sf)**
2. **Single, male (Sm)**
3. **Multiple (Mu)**
   Multiple authors, either collectively or in some form of dialogue. Blog posts with comments (even a single comment) count as *At:Mu*.
4. **Corporate (Co)**
   Any text which speaks in the name and interest of a company, association, party, syndicate, etc. A typical lexical indicator is the use of first person plural pronouns. In principle, even a text signed by a single author (president, CEO, etc.) could be *At:Co* if it is obvious that the text was not written and edited primarily by this person. This should, however, only be coded if the (collective) author clearly speaks in the name and interest of an organization (company, political party, etc.). Statutes, bylaws, regulations, and laws are always *At:Co*. Also, a project specification for a scientific project (grant application or description) is usually best coded as *At:Co*.

5. **Unknown (Un)**
   None of the above. Also, if the text was written by a single auhtor, but only initials are given (as is sometimes the case in newspapers), it is also *At:Un*.

2.3 Mode (Mo)

This category is interpreted as hierarchical in problematic cases.

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 some editing. There is no minimal length for *Mo:Wr*, even a Haiku is *Mo:Wr*.

2. **Spoken (Sp)**
   Any text that is clearly marked as spoken. Usually 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, probably without editing. Typically chat transcripts/ logs and unmoderated spontaneous discussions in forums.

4. **Blogmix (Bm)**
   Coders should assign this category to documents which contain some (prototypically initial) text which is *Mo:Wr*, but which is followed by significantly more spontaneous discussion of category *Mo:Qs*. Also, forum discussions with long and obviously edited posts in the middle should be *Mo:Bm*. Even smallest portions of *Mo:Qs* elements in a document should lead the coder to code *Mo:Bm*. Documents from *Mo:Bm* are probably almost always *At:Mu*

2.4 Audience (Au)

This category is interpreted as hierarchical in problematic cases. As a general rule, coders should prefer the "lower" level when in doubt.

1. **General (Ge)**
   If any educated native speaker (where "educated" is defined roughly in terms of 10 years of school such as in American High School or German Realschule) could summarize the text in her/his own words, it is *Au:Ge*. For example: a report about a new spacecraft written in an accessible style, recommendations for gardening, a short story, etc.

2. **Informed or Restricted (In)**
   If only a restricted group of native speakers (but not group defined by a certain profession) could adequately summarize the text, it is *Au:In*. For example: detailed discussion of some (computer) game (like a walk through or a secret guide), details about characters in some fictional book, etc. The same counts for discussion among owners of one specific phone, where a reader must know something about the phone to make sense of the discussion. Typically, in documents with *Au:In* there is a small amount of moderately specialized terminology.

3. **Professional (Pr)**
   A text which could only be adequately summarized by a person having a specific professional education. (It does not play a role whether there might be some people who have acquired the respective knowledge in their spare time. It suffices if there is a prototypical profession associated with the recipients of the text.) There is usually a higher amount of specialized terminology in such documents. For example: Texts obviously directed at plumbers, scientists, medical personnel, etc.

2.5 Aim (Ai)

This category is interpreted as hierarchical in problematic cases.

1. **Recommendation (Re)**
   Any text which advertises a certain product or service **in the interest of the author(s)**. The recommendation must be motivated by: commercial, religious, or political interest/belief or legal requirements. Not *Ai:Re*, however, is any review of a product (like a book or a movie) written by an independent person; this is *Ai:Di*. Also, vague recommendations as to how one should live her/his life, etc. are rather *Ai:Is* or even *Ai:Di* (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, etc. In general, the advice should be clear in a "do this, do that" sense; vague esoteric advice on how to achieve enlightenment or the like is *Ai:Di*. The advisor should be the original author of the document; a forum thread where someone asks for advice and possible solutions are discussed is *Ai:Di*, not *Ai:Is*. Also, if some (e. g., legal) situation is explained, and instructions can – in principle – be infered from the explanation, coders should code *Ai:If*. 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 *Ai:If* instead.

3. **Information (If)**
   Any text that provides statements of facts without any personal involvement or statement by the author(s). For example: science reports, statistics 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 no longer *Ai:If* but *Ai:Di*. All teaching materials are *Ai:Is*, not *Ai:If*. A grant application is *Ai:If* 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 *Ai:Di*. If a blog post is followed by even a single comment, it automatically becomes *Ai:Di*, even if the original post was rather *Ai:Is* or *Ai:If*. Forum threads are almost always *Ai:Di*.

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

2.6 Domain (Do)

This category is interpreted as hierarchical in problematic cases.

1. **Science (Sc)**
   Anything related to natural, social, and behavioral sciences (including linguistics). It does not have to be on an academic level. Notice that reports about education (schools and universities as institutions, job education) are by definition *Do:Po*.

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 *Do:Te*. If the text is about fundamental science related to some technology (i. e., not applied), it is *Do:Sc*, however.

3. **Medical (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 etc. These are *Do:Be*.

4. **Politics, Society, History (Po)**
   Anything related to politics, society or history. This includes any information and discussion about education, universities, the job market, international relations, war, etc.

5. **Business and Law (Bl)**
   Anything related primarily to business or legal matters. This includes discussion of personal economic or legal affairs. For example also discussion of economic problems connected with personal plans of emigration to another country. Statutes, bylaws, regulations, and laws are always *Do:Bl*.

6. **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 does not play a role. Includes literary science but not fiction itself (which is *Do:Ll*).

7. **Beliefs (Be)**
   Anything related to philosophy, 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 any medical method not accepted by the general medical profession (including borderline cases like hypnosis), code *Do:Be* instead of *Do:Me* should be assigned.

8. **Life and Leisure (Ll)**
   Any subject that is not covered by the above and that is a less substantial topic of everyday life. All fiction, talk about the (entertainment) media (if not clearly *Do:Po* or anything of the above), fashion, sports, games, pets, hobbies, cookery, gossip, etc. Self-portraits and CVs or documents containing someone's personal views on a collection of diverse matters without clear focus are *Do:Ll*.

3 Special cases and examples

- An imprint is always: *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Bl*.
- Texts published (officially) by political parties are by default: *At:Co, Mo:Wr, Au:Ge, Ai:Di, Do:Po*.
- General advice which does not entail knowledge transfer is *Ai:Di*, not *Ai:Is*. An example is advice on personal style or suggestions what to give away as a present.
- Blog entries by MPs or other active politicians are *At:Co* only if there is significant suspicion that the posts are written by a ghost writer, otherwise it is *At:Sf* or *At:Sm*.
- Job offers are most likely *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Bl*.
- Descriptions of study programs at universities are *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Sc* by definition.
- If the obvious aim of a text to get readers to donate something (money, tokens, bone marrow), give blood, etc., the text is *Ai:Re*.