COWCat text categorization guidelines 2013 (preliminary version, under constant revision)

Roland Schäfer
roland.schaefer@fu-berlin.de
German Grammar, Freie Universität Berlin

March 24, 2013

The following people contributed to the classification scheme and the CatTLE reference corpus:
Felix Bildhauer, Sarah Dietzfelbinger, Lea Helmers, Theresia Lehner, 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://hpsg.fu-berlin.de/cow/

There will be a corpus of German Web documents classified according to COWCat 2013 (CatTLE.de):

http://hpsg.fu-berlin.de/cow/cattle/

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 a value in each category (incl. at least one value for *Domain*). There must not be any blank or N/A fields.
- For each category, two authorities are selected among the coders. They are those who achieve the highest inter-rater agreement after 100 documents.
- In case the coder has significant difficulty assigning some value for a specific document, s/he should contact the two authorities and discuss it. This should not happen significantly more often than once in a hundred documents.

3 The taxonomy

3.1 Hierarchically interpreted categories

Some categories should be interpreted hierarchically in case of problematic decisions. I. e., if a coder cannot decide for *Mo*, *Au* or *Ai* which value to code, s/he should code the value (of the multiple applicable ones) which is highest in the category list.

3.2 Authorship (At)

To determine the value for Authorship (especially *At:Sf* and *At:Sm*), 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 *At:Mu*. A forum thread with only one poster is not *At:Mu*.

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 *At:Co* unless the text is about private matters (personal illness, own family, etc.). All 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.

3.3 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 at least 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, 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 *Mo:Qs*. Such documents usually involve

some form of dialogue. Typically, they are chat transcripts/logs and unmoderated spontaneous discussions in forums.

4. **Blogmix (Bm)**
   Coders should assign this category to documents which contain some strictly initial text which is *Mo:Wr*, but which is followed by significantly more spontaneous discussion of category *Mo:Qs*. The discussion typically focusses on the subject of the initial written part (as in blog posts or newspaper articles with comments). Coders should consider documents which consist of an estimated 25% or more of discussion as *Mo:Bm* instead of *Mo:Wr*. At the other end, more than roughly 75% of discussion make a document *Mo:Qs* instead of *Mo:Bm*. Documents from *Mo:Bm* are probably almost always *At:Mu*.

3.4 Audience (Au)

This category is interpreted as hierarchical in problematic cases. As a general rule, coders should prefer *Ge* when in doubt, because we consider it more important to find few truly non-general documents than a lot of which many are borderline cases. This category is not about readability per se. This means that syntactic complexity, stylistic well-formedness, complex document structure, etc. do not play a role. It is about the contents of the document, and in a limited way about the terminology.

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, straightforward recommendations for gardening, an simple short story, plainly written blog posts about the author's daily life, 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, a report about a snooker match, etc. The same counts for discussion among owners of (for example) some 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 same 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.

3.5 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 may be motivated by, for example: 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, spiritual instructions on how to achieve enlightenment, etc. 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 inferred indirectly 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. Linguistically, imperatives or similar constructions (*do this, do that,...*) are good indicators of *Ai:Is*. Statements and conditionals are indicators of *Ai:If*. (These linguistic generalizations are just hints and should be taken with a grain of salt.)

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 no longer *Ai:If* but *Ai:Di*. All teaching materials are *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 comments (more than roughly 25% of the whole document), it is *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 fictional text, including poems.

3.6 Domain (Do)

Each coder has to assign at least one domain, but may maximally assign four. The order of the assignments is not taken to imply a hierarchy or importance. This category is open, such that new domains can be added after tune-in coding.

1. **Science (Sc)**
   Anything related to empirical or "formal" 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 (cf. *Audience* for this). Notice that reports about education (schools and universities as institutions, job education) are not automatically *Do:Sc* if they are not about actual research. If they are more about organizational, political and institutional matters, they but might be, for example, *Do:Po*. *Do:Sc* does not include

   - Medical Science (cf. *Do:Me*)
   - Political Science, Sociology, Macroeconomics (cf. *Do:Po*)
   - Historical Science (cf. *Do:Hi*)
   - Microeconomics (cf. *Do:Bi*)
   - Jurisprudence (cf. *Do:Lw*)
   - Literary Science, Science of Art, Musicology (cf. *Do:Ar*)
   - Philosophy (cf. *Do:Ph*)
   - Theology (cf. *Do:Be*)

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, homeopathy, etc. These are *Do:Be*.

4. **Public Life and Infrastructure (Pi)**
   Anything related to the organization of public life, official authorities (fire brigade, police, military, public departments, schools, etc.) or infrastructure without larger social or political impact. Discussions about road construction, practical advantages of fire drills at schools, train delays, personal problems with job agencies. This also includes administrative information by schools and universities (organization of study programs, how to find an internship, etc.). As soon as there is any kind of political impact (Stuttgart 21, BER airport, general grievance over the job agency policies), the document is *Do:Po*, however.

5. **Politics, Society (Po)**
   Anything related to politics or society, but not under a primarily historical perspective. This includes any political discussion about education, universities, the job market, international relations, war, royalty, etc. Mere administrative information on/discussion about schools and universities, job agencies, etc. are *Do:Pi*, 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 *Do:Po*. Environmental issues under a political perspective are *Do:Po*.

6. **History (Hi)**
   Anything related to history. This is much like *Do:Po* 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.

7. **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.

8. **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 *Do:Po*, not *Do:Lw*. Once it has been officially ratified and there is discussion about its interpretations, a report about the interpretation is *Do:Po*. Statutes, bylaws, regulations, and laws are always *Do:Lw*.

9. **Fine 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. Bob Ross and Justin Bieber are also *Do:Ar*. Includes literary science but not fiction itself (which is *Do:Ll*). Games (including computer games) are *Do:Ll* instead.

10. **Philosophy (Ph)**
    Anything related to Philosophy. Philosophy is not empirical (like *Do:Sc*), and it is not bound by a specific religious view of the world (like *Do:Be*).

11. **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.

12. **Life and Leisure (Ll)**
    Any 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.

13. **Individuals (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 *Do:Iv*.

4 Special cases and examples (needs revision)

- Imprints and Terms and Conditions are always: *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Lw*.
- 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 but is rather the expression of a view 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 by companies are most likely *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Bi*. If they are offered by a public job agency, they are probably *Do:Pi*.
- Descriptions of study programs at universities or any administrative information regarding schools and universities are *At:Co, Mo:Wr, Au:Ge, Ai:If, Do:Pi* by definition. If there is a greater amount of scientific/scholarly content, then they might rather be *Au:Pr* and (maybe additionally) *Do:Sc*,but this is very unlikely.
- If the obvious aim of a text is to get readers to donate something (like money, tokens, bone marrow), give blood, etc., the text is *Ai:Re*.
- Biographies, autobiographies, and CVs are *Ai:If* and *Do:Iv* and most likely *Mo:Wr, Au:Ge*.