Report on the project discussion in November–December 2018

Olga Sozinova

January 2019

In order to create a corpus with data encompassing texts in 100 languages, we need to agree on useful meta information and on a sampling procedure. One of the most important meta feature is genre, because it has a direct effect on sampling (i.e. sampling from novels and technical texts is not the same). This report describes our decisions on genres that we will use in the project, how they correspond to the online sources observed by now, how the sampling strategies will differ according to different genres and what structure will our files and folders have.

1 Genres

We started to work on genres by consulting the book ”Variation Across Speech and Writing” by Douglas Biber (Biber 1991). We consider this source reliable, because Biber identifies genres empirically based on different linguistic characteristics. He lists 23 genres: Press Reportage, Press Editorials, Press Reviews, Religion, Popular Lore, Biographies, Humor, General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction, Face-to-face Conversations, Telephone Conversations, Interviews, Broadcasts, Spontaneous Speeches, Prepared Speeches, Personal Letters, Hobbies, Official Documents, Academic Prose, Professional Letters.

Since many of the languages that we include into analysis are endangered, we decided to include two more fine-grained categories: Oral Tradition and Written Tradition. These categories will serve for texts collected in linguistic expeditions and other language documentation activities (low resource languages).

However, fine-grained genres do not suit our analysis well. We need broader categories, which reflect difference in sampling procedure and which allow to include any not yet defined fine-grained category. We agreed on 6 broader categories: Fiction, Non-Fiction, Conversation, Professional, Technical, Grammar Examples.

Category *Fiction* is straightforward and includes different sub-types of fiction. *Non-Fiction* category is the broadest one: there are all types of journalistic, religious and not strictly structured texts (such as Personal Letters, Humor, Oral and Written Tradition). *Conversation* includes scripted conversations and interviews. Category *Professional* encompasses texts which involve any type of terminology, i.e. words of non-common knowledge (genre Hobbies also belongs here, because there is specific terminology as well). *Technical* texts are localization files (e.g. for Ubuntu), manuals, documentation to software and instructions to medicaments. We added a special category *Grammar Examples* for sentences and texts collected from grammar books. Mode of such sentences can be both spoken (scripted speech) and written (constructed sentences).

In addition to these broader categories, we are going to add a category *Random Sentences* on the final web source. It will be implemented as a function, that shows a sample of random sentences from different genres (or a specific genre). The sample will be drawn dynamically on the server side.

Table 1 shows the correspondence between broader categories, fine-grained genres, mode and online sources as examples.

2 Sampling

Since the data available is excessively heavy (more than 300GB only for the Wikipedia dumps), we cannot store every document on our server. We are going to create samples from the documents containing more than 10 000 numbers (the number of tokens can be changed in the future).

I propose three different sampling strategies: one for the documents falling into broad categories *Fiction*, *Non-Fiction*, *Professional*; another one for the documents from the *Conversation* category; and the third strategy for the *Technical* texts.

Sampling algorithm is the same for all categories. First, we define the number of samples, for example, three. During each sample, we draw a certain amount of tokens from the document, e.g. 10 000.

Table 1. Correspondence between genres, modes, sources

Mode	Genre (broad)	Genre (narrow)	Source example
written	Fiction	General Fiction	OPUS: Books
written	Fiction	Mystery Fiction	-
written	Fiction	Science Fiction	-
written	Fiction	Adventure Fiction	-
written	Fiction	Romantic Fiction	-
written	Non-Fiction	Press Reportage	OPUS: GlobalVoices
written	Non-Fiction	Press Editorials	-
written	Non-Fiction	Press Reviews	OPUS: NewsCommentary
written	Non-Fiction	Religion	Bible Parallel Corpus
written	Non-Fiction	Popular Lore	Wikipedia Dumps
written	Non-Fiction	Biographies	-
written	Non-Fiction	Humor	-
written	Non-Fiction	Prepared Speeches	OPUS: OpenSubtitles2018
written	Non-Fiction	Broadcasts	-
spoken	Non-Fiction	Oral Tradition	-
written	Non-Fiction	Written Tradition	-
written	Non-Fiction	Personal Letters	-
spoken	Conversation	Face-to-face Conversations	-
spoken	Conversation	Telephone Conversations	-
spoken	Conversation	Interviews	-
spoken	Conversation	Spontaneous Speeches	SketchEngine: CHILDES
written	Professional	Hobbies	-
written	Professional	Official Documents	SketchEngine: Eur-Lex
written	Professional	Academic Prose	-
written	Professional	Professional Letters	-
written	Technical	-	OPUS: Ubuntu
spoken/written	Grammar Examples	-	-
spoken/written	Random Sentences	-	-

Second, we find a random starting point for the current sample. The starting point should be always a beginning of a sentence. In case of *Fiction*, *Non-Fiction* and *Professional* texts, the best starting point would be at the beginning of a new logical part of the text. Such starting points can be searched by keywords, such as *chapter*, *section*, *article*. In the *Conversation* category, the starting point should not, when possible, break the natural flow of the conversation, e.g. if there is a question followed by an answer, the question should serve as a beginning of a sample. There is no specific instruction for starting points in the *Technical* category. Only if the text is divided into paragraphs or sections, the starting point should be at the beginning of a paragraph (e.g. in case of instructions to medicaments).

Third step is to extract the needed amount of tokens starting at the chosen point, save it and delete it from further sampling.

The total amount of tokens that we want to reach for each language is 50 000. Completely repeated documents will be omitted.

3 File structure

We agreed to store both raw *.txt and *.xml files for each sample. The names of the files will correspond to the following scheme: three letter ISO code followed by an underscore, followed by a three letter code representing genre, followed by an underscore and a digit running number. For example, eng_fic_1.txt.

Table 2 shows the correspondence between broader categories and the three letter codes I propose:

Table 2. Correspondence between broader categories and codes

Broader category	Code
Fiction	fic
Non-Fiction	nfi
Conversation	con
Professional	pro
Technical	tec
Grammar Examples	gre

XML documents will consist of a header containing meta information and a text. Meta information will include the following fields:

- language_name_wals: language name in the WALS 100 language sample (see langInfo_100LC.csv)
- language_name_glotto: language name in Glottolog 3.3 (see langInfo_100LC.csv)
- iso: ISO_639-3 code (see langInfo_100LC.csv)
- year_composed: in which year was the text written/recorded (if unknown put 'NA')

- year_published: which year was it published (if unknown put ’NA’)
- mode: spoken/written (if the description of a text or collection gives clear reference to a narrator, e.g. *was told by*, then it is ’spoken’)
- genre_(broad): either of the broad genres defined in Table 1
- genre_(narrow): either of the narrow genres defined in Table 1
- writing_system: give the ISO 15924 four letter code identifying the script used in the text (e.g. Latin: Latn, Cyrillic: Cyrl, etc.). See https://en.wikipedia.org/wiki/ISO_15924 for the full list of codes.
- special_characters: If there are particular characters/diacritics introduced in a text, give the reference to the section in the respective article/book/online source (phonetic key) where a description can be found
- short_description: short description of the text’s content (e.g. an English title given to oral stories)
- source: URL (with date) for online texts; bibliographic reference for books, etc.
- copyright_short: some sources give specific short copyright phrases that should be used (see example in PBC_example_eng)
- copyright_long: full copyright statement as given by the source (see example in PBC_example_eng)
- sample_type: ’whole’ (for the documents containing less or equal to 10 000 tokens) or ’part’ (for the samples taken from a larger document)
- comments: further comments that are necessary for understanding the transcriptions of texts

Each sentence of the text will be put inside of a tag containing an attribute ’sentence_id’ with a digit running number. This attribute can help us to find parallel texts, links to which could be further added into the meta information. Grammar examples should have three layers of tags: one for the actual sentence, one for the glosses, and one more for translation (English, Spanish or another language, if available).

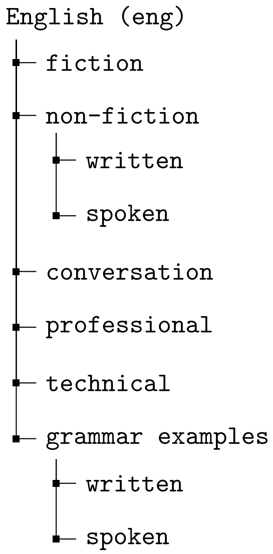

Figure 1. Example of the folder tree

4 Folder structure

The corpus folder will consist of the language folders, named by the language names and their ISO 639-3 codes. The first level folder names are stored in the file langInfo_100LC.csv, column folder_name. The second level of folders will include broader categories' names. Since there are only two broader categories that involve both spoken and written modes, namely *Non-Fiction* and *Grammar Examples*, there will be a third level in the folder structure for these two categories.

Figure 1 shows the intended folder structure for the corpus files.

References

Biber, Douglas (1991). *Variation across speech and writing.* Cambridge University Press.