About the Corpora

The corpora are collected from publicly available sources by a web crawler. The crawler checks for language, so as to mainly get texts consisting of the desired language*.

Each entry is tagged with it's date of publication. Where user comments are included they will be tagged with the date of the main entry.

Each entry is tagged with the type of entry, based on the type of website it is collected from (e.g. newspaper or personal blog) If possible, each entry is tagged with one or more subjects based on the title or keywords of the entry (e.g. if the entry comes from the sports section of a newspaper it will be tagged with "sports" subject).In many cases it's not feasible to tag the entries (for example, it's not really practical to tag each individual Twitter entry, though I've got some ideas which might be implemented in the future) or no subject is found by the automated process, in which case the entry is tagged with a '0'.

To save space, the subject and type is given as a numerical code.

Once the raw corpus has been collected, it is parsed further, to remove duplicate entries and split into individual lines. Approximately 50% of each entry is then deleted. Since you cannot fully recreate any entries, the entries are anonymised and this is a non-profit venture I believe that it would fall under Fair Use.

Corpus Sample

tagesspiegel.de 2010/12/03 1 7 Er ist weder ein Abzocker noch ein Ausbeuter, er ist kein Betrger, er haut niemanden in die Pfanne oder betrgt ihn um seinen gerechten Anteil, er steht zu seinem Wort und erfllt seine Vertrge sinngem und feilscht nicht wegen irgendwelcher Lcken im Maschendraht des Kleingedruckten der Vertrge.spiegel.de 2010/11/30 1 1,6 Diplomaten sehen Clintons Direktive als Besttigung einer alten Regel: Diezeit.de 2009/10/22 1 2,10 Warum schaffen wir nicht eine Whrung, die diese Aufgaben erfllt anstatt den Forderungen der Geldwirtschaft hinterherzuhecheln, die niemals erfllt werden knnen?

* You may still find lines of entirely different languages in the corpus. There are 2 main reasons for that:1. Similar languages. Some languages are very similar, and the automatic language checker could therefore erroneously accept the foreign language text.2. "Embedded" foreign languages. While a text may be mainly in the desired language there may be parts of it in another language. Since the text is then split up into individual lines, it is possible to see entire lines written in a foreign language.Whereas number 1 is just an out-and-out error, I think number 2 is actually desirable, as it will give a picture of when foreign language is used within the main language.