COPPA V2.0: Corpus Of Parallel Patent Applications

The segments included in the corpus are obtained by aligning the sentences of the abstracts and titles of published PCT applications with their translations, the translations having been produced by professional patent translators (More than 200,000 new PCT applications are published every year). It is therefore a gold mine for linguistic research such as terminology extraction, translation memory building and research on Machine Translation.

With the goal of supporting innovation in the Machine Translation field, WIPO offers the updated corpus under the same conditions as before, the product being notably free of charge for academic and private research institutions for research purposes only (without redistribution right); in return those institutions commit to share their published results with WIPO.

WIPO hopes that the wide availability of this improved corpus will actively contribute to progress in building more accurate machine translation systems for patent texts with the ultimate goal of lowering the linguistic barrier for inventors and the general public and of improving the efficiency and the accessibility of the international patent system.

1 Statistics

The corpus now contains more than 300 Million words (English-French), for comparison (only for English-French), the previous COPPA version contained 180 Million words. See Table 1 for full details:

Language pair	Documents	Sentences	Words	Characters
en-de	289'287	982'510	36'814'520	225'972'826
en-es	18'303	62'057	2'328'713	14'624'745
en-fr	2'570'292	10'557'032	316'271'950	2'006'750'520
en-ja	312'664	1'036'614	42'127'479	264'578'974
en-ko	41'093	120'534	5'813'474	37'047'347
en-pt	2'001	7'000	261'843	1'696'039
en-ru	6'972	37'261	1'241'791	7'841'040
en-zh	83'359	195'317	7'325'443	47'401'578
Total	3'323'971	12'998'325	412'185'213	2'605'913'069

Table 1: Statistics for the complete corpus. The total does not reflect unique documents as all the documents are available in English and French (a Japanese document - in the en-ja corpus - will also be part of the en-fr subcorpus)

2 Organization and file structure

The corpus is organized in two data formats:

1. XML documents adhering to the TEI standard, with one file per document and one link file per document pair;
2. Plain text files in the Moses format, containing all data per language pair for easy machine translation training.

2.1 TEI Documents

TEI documents are located in the "tei" folder, one tar-archive per language. An archive can be "untared" with the following command:

```
tar -xzf CoppaV2.en.tgz
```

Which will result in a folder structure as depicted below:

```
en
├── WO1978
│   └── 00
│       └── 00
│           ├── WO1978000001.xml
│           ├── WO1978000002.xml
│           ├── WO1978000003.xml
│           ├── WO1978000004.xml
│           ├── WO1978000005.xml
.           .
.           .
.           .
├── WO1979
│   └── 00
│       ├── 00
│       │   ├── WO1979000001.xml
│       │   ├── WO1979000002.xml
│       │   ├── WO1979000003.xml
│       │   ├── WO1979000004.xml
│       │   ├── WO1979000005.xml
│       │   ├── WO1979000006.xml
.       .   .
.       .   .
.       .   .
```

Documents are organized primarily by year, with two levels of subfolders based fragments of document ids. An example document has the following (shortened) structure:

```
<?xml version="1.0" encoding="utf-8"?>
<TEI.2 id="WO1978000001-en" lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>LOW TEMPERATURE SYNTHESIS OF VITREOUS BODIES AND THEIR
        INTERMEDIATES</title>
      </titleStmt>
      <publicationStmt>
         <availability>
           <p>COPYRIGHT WIPO 2016, for questions contact
           patentscope@wipo.int</p>
         </availability>
      </publicationStmt>
    </fileDesc>
    <notesStmt>
      <note type="ID">WO1978000001</note>
      <note type="AD">19780602</note>
      <note type="ANID">US1978000002</note>
      <note type="DP">19781020</note>
      <note type="IC">B01D 17/02;</note>
      <note type="LGF">EN</note>
      <note type="OF">WO</note>
    </notesStmt>
  </teiHeader>
  <text>
    <body>
      <head id="1" lang="en">LOW TEMPERATURE SYNTHESIS OF VITREOUS BODIES
      AND THEIR INTERMEDIATES</head>
      <div type="abstract">
        <p id="2">
          <s id="2:1" lang="en">A method of making glass of high purity
          and in virtually unlimited shapes via solution deposition on
          a porous self-supporting body by reaction between a first
          solution and a second solution; an a product made thereby.</s>
          ...
```

```
        </p>
      </div>
    </body>
  </text>
</TEI.2>
```

2.2 Meta data information

Each patent application contains:

1. the ID (Any document is available on WIPO search engine PATENTSCOPE using: `http://patentscope.wipo.int/en/ID`)
2. the AD: application date, in format YYYMMDD
3. the DP: date of publication, in format YYYMMDD
4. IC: the IPC code - International Patent Classification - (eg. for IPC class "B01D 17/02" consult `http://www.wipo.int/ipcpub/?symbol=B01D0017020000`)
5. LGF: Language of filling, the original language of the patent application (useful to know the original human translation direction)
6. OF: the patent office code (see `http://www.wipo.int/standards/en/pdf/03-03-01.pdf`)

2.3 Link files

Link files are stored in tei/CoppaV2.links.tgz for all language pairs with one folder per language pair. A similar command as above:

```
tar -xzf CoppaV2.links.tgz
```

will result in the following folder structure:

```
links
├── de_en
│   ├── WO1986
│   │   └── 00
│   │       └── 71
│   │           └── WO1986007149.lnk
│   ├── WO1989
│   │   └── 00
│   │       ├── 14
│   │       │   └── WO1989001410.lnk
│   │       ├── 41
│   │       │   └── WO1989004114.lnk
│   │       ├── 45
│   │       │   └── WO1989004577.lnk
│   │       ├── 59
│   │       │   └── WO1989005944.lnk
│   │       └── 65
│   │           └── WO1989006509.lnk
│   ├── WO1990
│   │   ├── 00
│   │   │   ├── 00
│   │   │   │   ├── WO1990000011.lnk
│   │   │   │   ├── WO1990000024.lnk
.   .   .   .   .
.   .   .   .   .
.   .   .   .   .
```

Link files contain sentence alignment information for document pairs with the same ids for each language pair, for instance the file `links/de_en/WO1986/00/71/WO1986007149.lnk` looks like this:

```
<linkGrp fromDoc="Xml/de/WO1986/00/71/WO1986007149.xml"
toDoc="Xml/en/WO1986/00/71/WO1986007149.xml" score="0.540672">
  <link type="1-1" xtargets="1;1" score="1" />
```

```
  <link type="1-1" xtargets="2:1;2:1" score="0.388821" />
  <link type="1-1" xtargets="2:2;2:2" score="0.712389" />
  <link type="1-1" xtargets="2:3;2:3" score="0.446505" />
  <link type="1-1" xtargets="2:4;2:4" score="0.73225" />
  <link type="1-1" xtargets="2:5;2:5" score="0.627393" />
  <link type="1-1" xtargets="2:6;2:6" score="0.463875" />
  <link type="1-1" xtargets="2:7;2:7" score="0.41347" />
</linkGrp>
```

It references the documents which are aligned by this link file and the overall score for the alignment. Each entry identifies the paragraph or sentences id (divided by ";") and the sentence-level alignment quality. Different alignment types up to 4-4 alignments are possible.

2.4 Plain-text moses format

Plain-text files are located in the "moses" folder, with one folder per language pair.

```
moses
├── de_en
│   ├── CoppaV2.de.gz
│   ├── CoppaV2.en.gz
│   └── CoppaV2.meta.gz
├── es_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.es.gz
│   └── CoppaV2.meta.gz
├── fr_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.fr.gz
│   └── CoppaV2.meta.gz
├── ja_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.ja.gz
│   └── CoppaV2.meta.gz
├── ko_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.ko.gz
│   └── CoppaV2.meta.gz
├── pt_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.meta.gz
│   └── CoppaV2.pt.gz
├── ru_en
│   ├── CoppaV2.en.gz
│   ├── CoppaV2.meta.gz
│   └── CoppaV2.ru.gz
└── zh_en
    ├── CoppaV2.en.gz
    ├── CoppaV2.meta.gz
    └── CoppaV2.zh.gz
```

Each file is a compressed single utf-8 plain-text file, lines correspond to each other for each language pair, but not across language pairs. For each pair a meta data file is included which informs about the source document id for each line, the alignment type, and the link quality.

```
WO1986007149    1-1     1
WO1986007149    1-1     0.388821
WO1986007149    1-1     0.712389
WO1986007149    1-1     0.446505
```

Based in this information it is also possible to find the original XML TEI document. For the previous example, you can access the TEI files in:

- `Xml/en/WO1986/00/71/WO1986007149.xml`
- `Xml/fr/WO1986/00/71/WO1986007149.xml`

The source-target sentence links are available in Xml link files, for the previous example, the two language-pairs link file is:

- `Xml/links/fr_en/WO1986/00/71/WO1986007149.lnk`

Note that the original patent application is available on-line on our search engine PATENTSCOPE, at `https://patentscope.wipo.int/search/en/WO1986007149`. Other documents can be looked up analogously.

3 Final remarks

For scientific publications, please cite this reference:

Junczys-Dowmunt, M., Pouliquen, B., and Mazenc, C. (2016). COPPA V2.0: Corpus Of Parallel Patent Applications; Building Large Parallel Corpora with GNU Make. Language Resources and Evaluation (LREC'16). Portorož, Slovenia.

For questions, please contact patentscope@wipo.int