

WIPO Corpus of Parallel Patent Applications

Technical details

We have chosen to comply with the TMX format, preliminary sample corpus and statistics are available, but this may change in the future. Each document contains title and abstracts available in both languages. The first set of files (PCT2010-en-fr.tmx) contain one "translation unit" for the title and the abstract. The second file is the result of our processing for tokenizing, segmenting and aligning the abstract and maybe more suitable as the translation units are shorter than 80 words.

Sample data

Two sample files are available, each containing a selection of patent applications published in 2010:

- [samplePCT2010-en-fr.tmx](http://www.wipo.int/patentscope/translate/coppa/samplePCT2010-en-fr.tmx.zip) (containing title and abstracts as published on our search engine [PATENTSCOPE](#)), available for download at the following address: <http://www.wipo.int/patentscope/translate/coppa/samplePCT2010-en-fr.tmx.zip>
- [samplePCT-seg-2010-en-fr.tmx](http://www.wipo.int/patentscope/translate/coppa/samplePCT-seg-2010-en-fr.tmx.zip) (the corresponding segmented version) , available at the following address: <http://www.wipo.int/patentscope/translate/coppa/samplePCT-seg-2010-en-fr.tmx.zip>

File	Size (bytes)	Size (compressed)	N° translation units	N° documents	N° characters	N° words
samplePCT-seg-2010-en-fr.tmx.gz	1526934	262466	2670	500	346592	55840
samplePCT2010-en-fr.tmx.gz	1231384	313542	1000	500	417608	67250

Format overview:

The left part shows two translation units taken from the first file, the right part shows the corresponding first translation units.

Each translation unit contains:

- (1) the ID (Any document is available on WIPO search engine PATENTSCOPE using: <http://www.wipo.int/patentscope/search/en/ID>)
- (2) the applicant name (2) the IPC code (eg. for IPC class "F01N 3/02" <http://www.wipo.int/ipcpub/?symbol=F01N0003020000>)
- (3) the type of segment: title or abstract

Note that in the segmented version only the main IPC code is retained, the "srclang" attribute indicate in which language the application was filed.

```
- <tu tuid="WO2010100270" srclang="EN">
  <prop type="Att.applicant">MICRONIC MYDATA AB</prop>
  <prop type="Att.IPC">G03F 7/20</prop>
  <prop type="Txt.DocType">PCT Title</prop>
  - <tuv xml:lang="FR">
    <seg>PROCÉDÉ ET APPAREIL POUR UN ÉCLAIRAGE STATISTIQUE</seg>
  </tuv>
  - <tuv xml:lang="EN">
    <seg>METHOD AND APPARATUS FOR STATISTICAL ILLUMINATION</seg>
  </tuv>
</tu>
- <tu tuid="WO2010100270">
  <prop type="Att.applicant">MICRONIC MYDATA AB</prop>
  <prop type="Att.IPC">G03F 7/20</prop>
  <prop type="Txt.DocType">PCT Abstract</prop>
  - <tuv xml:lang="en">
    - <seg>
      The technology disclosed relates to an illumination source including numerous laser diodes. In particular, it relates to extending the duty cycle and/or reducing the frequency of component replacement by detecting failure of one or more individual laser diodes and compensating for the failure, without replacing the laser diodes. The technology disclosed can be used in cases of catastrophic laser diode failure by changing the power of remaining laser diodes to restore illumination to the coherence function similar to the pre-failure illumination field. Particular aspects of the technology disclosed are described in the claims, specification and drawings.
    </seg>
  </tuv>
  - <tuv xml:lang="fr">
    - <seg>
      La présente technologie porte sur une source d'éclairage comprenant de nombreuses diodes lasers. En particulier, elle porte sur l'extension du rapport cyclique et/ou la réduction de la fréquence de remplacement de composant par la détection de défaillance d'une ou de plusieurs diodes lasers individuelles et par la compensation de la défaillance, sans remplacement des diodes laser. La technologie divulguée peut être utilisée dans des cas de défaillance de diodes lasers catastrophique par un changement de la puissance des diodes lasers restantes pour restaurer un éclairage à la fonction de cohérence similaire au champ d'éclairage avant la défaillance. Des aspects particuliers de la technologie divulguée sont décrits dans les revendications, la description et les dessins.
    </seg>
  </tuv>
  - <tuv xml:lang="EN">
    - <seg>
      in particular, it relates to extending the duty cycle and/or reducing the frequency of component replacement by detecting failure of one or more individual laser diodes and compensating for the failure, without replacing the laser diodes
    </seg>
  </tuv>
</tu>
- <tu tuid="WO2010100270" srclang="EN">
  <prop type="Att.mainIPC">G03F 7/20</prop>
  <prop type="Att.applicant">MICRONIC MYDATA AB</prop>
  <prop type="Txt.DocType">PCT Abstract</prop>
  - <tuv xml:lang="FR">
    - <seg>
      en particulier, elle porte sur l'extension du rapport cyclique et/ou la réduction de la fréquence de remplacement de composant par la détection de défaillance d'une ou de plusieurs diodes lasers individuelles et par la compensation de la défaillance, sans remplacement des diodes laser
    </seg>
  </tuv>
  - <tuv xml:lang="EN">
    - <seg>
      in particular, it relates to extending the duty cycle and/or reducing the frequency of component replacement by detecting failure of one or more individual laser diodes and compensating for the failure, without replacing the laser diodes
    </seg>
  </tuv>
</tu>
```

Statistics on the full corpus

The of the corpus will be more than **170 Million words** (for comparison, the [JRC-acquis](#) is about 35 Million words for English French, [Europarl](#) is about 50 Million and [MultiUN](#) is 370 Million)

Statistics for the non-segmented corpus

The number of characters and number of words are computed on the English version.

File	Size (bytes)	Size (compressed)	N° translation units	N° documents	N° characters	N° words
PCT1990.en-fr.tmx.gz	36694158	10461698	32216	16108	12389179	2031516
PCT1991.en-fr.tmx.gz	45850598	13023362	40324	20162	15368004	2514227
PCT1992.en-fr.tmx.gz	52262323	14762465	45878	22939	17453328	2849210
PCT1993.en-fr.tmx.gz	59439901	16746272	52156	26078	19878533	3236790
PCT1994.en-fr.tmx.gz	68063335	19068673	59979	29990	22738036	3698661
PCT1995.en-fr.tmx.gz	81712229	23009749	71246	35623	27445208	4464104
PCT1996.en-fr.tmx.gz	97487566	27242507	84346	42173	32685712	5295939
PCT1997.en-fr.tmx.gz	116634423	32385886	100540	50270	39103930	6346316
PCT1998.en-fr.tmx.gz	138461725	38045901	118961	59482	46410250	7498410
PCT1999.en-fr.tmx.gz	157664656	42993064	135807	67904	52983073	8540739
PCT2000.en-fr.tmx.gz	183286264	49528480	159118	79560	61631144	9927569
PCT2001.en-fr.tmx.gz	225979493	60147602	196627	98314	75615213	12124179
PCT2002.en-fr.tmx.gz	235508828	62585278	206434	103219	78342857	12562245
PCT2003.en-fr.tmx.gz	242984159	64405559	214049	107026	80658209	12941821
PCT2004.en-fr.tmx.gz	259276905	68316587	228369	114185	86526417	13912787
PCT2005.en-fr.tmx.gz	283748405	74528628	249714	124857	94813830	15230009
PCT2006.en-fr.tmx.gz	315233289	83724088	276448	138224	105385771	16947646
PCT2007.en-fr.tmx.gz	340432581	89744143	299084	149542	113648984	18280509
PCT2008.en-fr.tmx.gz	359409278	94079962	314856	157429	119965312	19308796
PCT2009.en-fr.tmx.gz	366210843	95180375	316882	158441	122485532	19739503
PCT2010.en-fr.tmx.gz	356076811	91367007	303472	151736	119700660	19272626
TOTAL	4022417770	1071347286	3506506	1753262	1345229182	216723602

Statistics for the segmented corpus

File	Size (bytes)	Size (compressed)	N° translation units	N° documents	N° characters	N° words
PCT-seg-1990.en-fr.tmx.gz	46473037	9095220	81046	16055	10573333	1722269
PCT-seg-1991.en-fr.tmx.gz	57468664	11207227	100300	20085	12981914	2109601
PCT-seg-1992.en-fr.tmx.gz	65418799	12694222	114026	22847	14783102	2398012
PCT-seg-1993.en-fr.tmx.gz	74296322	14401080	129432	25968	16845464	2724730
PCT-seg-1994.en-fr.tmx.gz	84817778	16375984	148049	29872	19244159	3111219
PCT-seg-1995.en-fr.tmx.gz	100639428	19584649	174485	35478	23027411	3720341
PCT-seg-1996.en-fr.tmx.gz	120323625	23239275	208084	42012	27516077	4433081
PCT-seg-1997.en-fr.tmx.gz	142763728	27413781	246961	50021	32611383	5262382
PCT-seg-1998.en-fr.tmx.gz	169492474	32290339	291985	59186	38860504	6247812
PCT-seg-1999.en-fr.tmx.gz	192548288	36463180	330906	67545	44384533	7119477
PCT-seg-2000.en-fr.tmx.gz	223035175	41953713	383722	79146	51606674	8273773
PCT-seg-2001.en-fr.tmx.gz	274677688	50899788	472875	97814	63446267	10131499
PCT-seg-2002.en-fr.tmx.gz	282429325	52250045	488418	102495	64822714	10354019
PCT-seg-2003.en-fr.tmx.gz	291862364	53725017	506822	106349	66612714	10647279
PCT-seg-2004.en-fr.tmx.gz	312427052	57178666	541576	113456	71487472	11458800
PCT-seg-2005.en-fr.tmx.gz	347754061	63049665	604481	124079	79365357	12720470
PCT-seg-2006.en-fr.tmx.gz	392615206	71210918	685938	137533	89200185	14326117
PCT-seg-2007.en-fr.tmx.gz	426685344	76461897	749514	148770	96389566	15481121
PCT-seg-2008.en-fr.tmx.gz	458223761	80905756	811288	156985	102476871	16467721
PCT-seg-2009.en-fr.tmx.gz	470180933	81704247	838687	158076	104190300	16760333
PCT-seg-2010.en-fr.tmx.gz	451878162	77132105	808354	151345	99855142	16043201
TOTAL	4986011214	909236774	8716949	1745117	1130281142	181513257

Bibliography

- **Un corpora** (<http://www.uncorpora.org/>): Alexandre Rafalovitch & Robert Dale: United Nations general assembly resolutions: a six-language parallel corpus. MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada; pp.292-299.
- **MultiUN** (<http://www.euromatrixplus.net/multi-un/>): Andreas Eisele & Yu Chen: MultiUN: a multilingual corpus from United Nation documents. LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta; pp.2868-2872.
- **JRC-acquis** (<http://langtech.jrc.it/JRC-Acquis.html>): Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, & Daniel Varga: The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings, Genoa, Italy, 22-28 May 2006; pp.2142-2147

- **Europarl** (<http://www.statmt.org/europarl/>): Philipp Koehn: Europarl: a parallel corpus for statistical machine translation. MT Summit X, Phuket, Thailand, September 13-15, 2005, Conference Proceedings: the tenth Machine Translation Summit; pp.79-86. [Philipp Koehn: Europarl: a parallel corpus for statistical machine translation. MT Summit X, Phuket, Thailand, September 13-15, 2005, Conference Proceedings: the tenth Machine Translation Summit; pp.79-86.
- **TAPTA** (<http://www.wipo.int/patentscope/translate/>): Bruno Pouliquen, Christophe Mazenc & Aldo Iorio: Tapta: a user-driven translation system for patent documents based on domain-aware statistical machine translation. [EAMT 2011]: proceedings of the 15th conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium; eds. Mikel L.Forcada, Heidi Depraetere, Vincent Vandeghinste; pp.5-12.