

# The Open-Content Text Corpus project

Piotr Bański

Institute of English Studies,  
University of Warsaw  
pkbanski@uw.edu.pl

Beata Wójtowicz

Dept. of African Languages and Cultures,  
University of Warsaw  
b.wojtowicz@uw.edu.pl

**We wish to acknowledge support from grant # N104 050437  
from the Polish Ministry of Science and Higher Education.**

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)

# OCTC.sourceforge.net: basic information

- text corpus and tools available under the GNU GPL
- hosted and distributed by SourceForge
- targetting under-resourced languages (the other ~30 are also welcome)
- monolingual + aligned parts, the latter created by pointing into the monolingual texts
- TEI XML, extending the architecture of the National Corpus of Polish
- designed for synergy with FreeDict.org and Apertium.org
- first texts: Universal Declaration of Human Rights (55 languages)
- more data coming
  - Google Summer of Code
  - LREC participants (would that include you?)