OPUS is a growing collection of parallel corpora for many languages and various domains. The collection becomes pretty big and includes a variety of data sets and tools that are not only useful for statistical machine translation. OPUS has been extended a lot since its first appearance in 2003. Actually the best birthday present would be if anyone would decide to start a mirror of OPUS. Let me know if you are interested.

Here some of the highlights:

- over 150 languages and language variants
- over 5 billion aligned translation units
- downloads in XML/XCES, plain text (Moses/SMT) and TMX
- raw, tokenized and machine-annotated data
- monolingual data sets (for language modeling)
- search interfaces

Some recent news and data sets:

- EUbookshop: a large but noisy corpus (converted from PDF)
- Tatoeba: a small but clean corpus with many languages
- OpenSubtitles2012: an improved version of the 2011 version
- coming soon: OpenSubtitles2013 - an extension of OpenSubtitles2012
- UN, MultiUN, Europarl v7: aligned for all language combinations
- word alignments and phrase tables for the majority of bitexts

The Web Site: http://opus.lingfil.uu.se
More information: http://opus.lingfil.uu.se/trac/wiki

[from Corpora-list]