This constitutes the new release of the LASSY Large corpus:
a large corpus of automatically syntactically annotated sentences
of Dutch.

version september 2023

- all sentences have been parsed again (spring/summer 2023)

- automatically added UD annotations (both in XML, and in separate
  conllu files)

- automatically added some further, new, attributes in XML
  (is_np, is_vorfeld, is_nachfeld)

- more meta-data
  -- for some corpora, <metadata>
  -- <parser build= date= cats= skips=>
  -- <sentence sentid=>

- for the larger corpora, the corpora are now split in sub-parts
of 5000 sentences (rather than 10000 sentences as in the previous
release). The parts can be easily aligned: if a part was called
xxxxxx_part01093 in the previous release, then the same sentences
can now be found in the parts xxxxxx_part01093a and xxxxxx_part01093b.

- The annotations are no longer distributed in a special ("COMPACT",
see below) format, but rather as zip files which contain xml files
(each xml file contains the annotation of a single sentence).

version April 2016

- all sentences have been parsed again (January-March 2016)
- we no longer use postag and lemma from the original Sonar500
  but we use Alpino output instead (much better quality)
- as a result, postag and lemma is now available for all sub-corpora

==========

FORMATS

==========

The directory SUITES contains the sentences of the corpus, in a format
that is suitable to Alpino.

The directory LOGS contains the log files of running the parser.

The directory UD contains the CONLLU files of the automatically derived
UD annotations (automatically derived on the basis of the Alpino parses)

==========

TREEBANK FORMATS

==========

Treebanks can come in one of the following formats:

1. XML files, one XML file for one sentence (in the distribution, these
are combined together in zipped files, *.zip)
2. COMPACT format: gzipped concatenations of XML files, using dictzip
3. DACT format: database format (Oracle's dbxml) as understood by tools
such as DACT. It is also the format the PaQu uses internally.

The COMPACT and DACT formats can be automatically derived from the
(zipped) XML files format.

==========

TREEBANK TOOLS

==========

There are various tool sets to work with corpora in the various formats.

1. The older XML tools, originally by Geert Kloosterman, with later
additions by Daniel de Kok and Peter Kleiweg, which are part
of the Alpino distribution. These tools include
   act   - various tools for converting treebanks in XML and COMPACT format
   miniact   - simplified version of act
   dtsearch - search treebanks in compact format
   dtview   - view treebanks in compact format
Links:
Alpino homepage: https://www.let.rug.nl/vannoord/alp/Alpino/
Alpino at github: https://github.com/rug-compling/Alpino

2. A set of tools to convert treebanks between the three different
formats by Daniel de Kok and colleagues. These are available at github
via: https://rug-compling.github.io/alpinocorpus/

3. simple GO program to convert ZIP formation into COMPACT, and
simple GO program to convert COMPACT into DACT format
Available in the sub-directory cmd.

4. PaQu web-application by Peter Kleiweg (funded by Clariah). Some of
the corpora of LassyLarge are already present there. In addition, you
can upload corpora in various formats, including parsed corpora in the
zip format of this distribution. https://paqu.let.rug.nl:8068/

5. Dact. Tool by Daniel de Kok to search and display treebanks in the
Dact format. https://rug-compling.github.io/dact/

==========

More info at https://www.let.rug.nl/vannoord/Lassy

==========

Gertjan van Noord
g.j.m.van.noord@rug.nl
Oktober 2010
July 2012
February 2016
September 2023