Lassy Small Release 6, April 2021

*** CHANGES wrt Lassy Small release 5, september 2018
many small corrections

+added morpheme boundaries in lemma for compounds, using '_' symbol
+added features as produced by the parser (automatically)
+added named entity information as available in SONAR1
+improved link with DutchSemCor senses
+improved meta-data information
+sentence id is added
+UniversalDependency info is added, both in XML as well as in
separate conllu files (directory CONLLU)

The bare XML files as manually edited are available in Treebank/
The XML files with all the above additional information added is
available in Enhanced/

The dact files are based on Enhanced/

*** Documentation

Part of speech and lemma annotation is documented in POS_manual.pdf; the file POS.txt contains all part-of-speech tags with typical examples; the file POSatts.txt illustrated how part of the part-of-speech tags are represented in the XML files by separate attributes.

The syntactic dependency annotations are documenten in sa-man_lassy.pdf

Meta-data is documented briefly in README-meta.txt

suites.txt is a list of all sub-parts of the LassySmall corpus.

*** What are these directories?

CONLU/

- CONLLU contains the UD annotations, automatically derived from the original Lassy Small annotations,
using "alud". https://github.com/rug-compling/alud

Enhanced/

- Enhanced contains both XML files and DACT files of the Lassy corpus, augmented with
--- metadata
--- Named Entity information, if available in Sonar-500
--- Sense ID information, if available in DutchSemCor

--- Additional attributes provided by the automatic Alpino parser
Automatically derived from Treebank/ and other resources.

FREQ/
- this directory contains various counts derived from the annotations in Treebank/

LP/
- this directory contains part-of-speech and lemma information in a tabular format for all sentences of Lassy Small.
Automatically derived from Treebank/

MWU/
- this directory contains all multi-word-units from Lassy Small, with part-of-speech and lemma information.

Suites/
- this directory contains the tokenized sentences with their key, both in text format and in Prolog format.

Treebank/
- this directory contains the XML files of the linguistic annotations of Lassy Small. These are the files that
are used for manual correction.

TRIPLES/
- frequency information of all dependency triples of Lassy Small